EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Testing Dynamic General Equilibrium Models

with application to calibrated and simulated business cycle models

Eva Ortega

Thesis submitted for assessment with a view to obtaining
the Degree of Doctor of the European University Institute

Florence, December 1997

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Testing Dynamic General Equilibrium Models

with application to calibrated and simulated business cycle models

Eva Ortega

The Thesis Committee consists of:

Prof. Fabio Canova, Universitat Pompeu Fabra Barcelona, co-supervisor
" Tryphon Kollintzas, University of Athens
" Alfonso Novales, Universidad Complutense de Madrid
" Mark Salmon, EUI and City University London, Supervisor

Completing a Ph.D. at the E.U.I. has been to me much more of a life experience than anything else. I want to thank all the people who have contributed to that intensive, tough and wonderful part of my life.

I want to thank especially the unconditional support of my family and the warmth of the invaluable friends I have found inside the Badia walls over these years and with whom I have shared unforgettable moments outside them. I would have never arrived to this point of putting an end to my Ph.D. if it were not for the patience, support, example and encouragement I have received from some friends of mine whom I have had the enormous luck to have as colleagues, especially Angel, Chiara, Humberto and Susana.

Finally, I want to express my gratitude to my supervisors, Fabio and Mark, for putting up with me and my doubts all these years, and, of course, to the understanding and efficient presence of Jacqueline, Jessica and Marcia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stochastic dynamic general equilibrium models have become in recent years the central paradigm for the analysis and understanding of macroeconomic fluctuations. As Gali (1995) puts it, though early applications to business cycle models (e.g., Kydland and Prescott (1982)) were generally restricted to model economies for which technology shocks were the only sources of fluctuations, and where built-in classical assumptions guaranteed the optimality of equilibrium allocations, the flexibility of that paradigm has been illustrated in a number of recent papers which have developed models of economies characterized by the presence of non-classical features (e.g., imperfect competition as in Rotemberg and Woodford (1991), Gali (1994) or Ubide (1995) and/or alternative sources of fluctuations (e.g., shocks to government spending as in Christiano and Eichenbaum (1992) or Baxter and King (1993)). These efforts to enrich the basic framework have been conducted with the objective of improving its empirical relevance and performance.

How to obtain and assess the quantitative implications of stochastic dynamic general equilibrium models gave rise to the development of the calibration methodology, which obtains quantitative predictions from a fully articulated, artificial model economy and compares them to the particular set of observed stylized facts the model wanted to help explain. However, classical pieces in the calibration literature (e.g., Kydland and Prescott (1982) or (1991)) are typically silent on the metric one should use to evaluate the quality of the approximation of the model to the data. The approach favored by most calibrators is to glare over the exact definition of the metric

used and informally assess the properties of simulated data by comparing them to the set of observed stylized facts.

This dissertation focusses on how to evaluate the success of a model in replicating the stylized facts it wanted to explain, and on how to compare the success of alternative model specifications. It contributes to the literature on testing dynamic general equilibrium models asking the question of whether formal approaches provide satisfactory alternatives to this informal model evaluation, and answering by (i) critically reviewing recent procedures suggested to assess the fit of calibrated models, (ii) proposing a new one and (iii) comparing the performance of alternative model evaluation methodologies when applied to standard Real Business Cycle (BBC) models.

In studying the issue of evaluating stochastic dynamic general equilibrium models, we have paid particular attention to two important methodological problems which are not exclusive of calibration, so that the methods and solutions reviewed and suggested in the following pages apply to and involve a wide range of economic and econometric problems. The first problem is the fact that most models do not have an exact analytical solution and therefore the stochastic dynamic equilibrium of the model has to be approximated. There has been been a large amount of effort devoted to finding accurate approximate solution methods using simulation techniques (see the January 1990 issue of the Journal of Business and Economic Statistics or Marcet (1994)).

The second problem is how to give values to the deep parameters of the model. This issue is extremely related to testing the model since parameter values are essentially chosen so as to obtain the best fit of the model. Standard econometric techniques are often not applicable because of the lack of observed data but mainly because they do not make sense from a calibrator's point of view. As stated in Kydland and Prescott (1991), "..., no attempt is made (in the Calibration approach) to determine the true model. All model economies are abstractions and are by definition false", and hence, as Pagan (1994) stresses, unlikely to obey the 'axiom of correct specification'. Standard econometric estimation criteria such as maximizing the likelihood function of the actual data given the model desing make the assumption that the model is the true DGP of the actual data, which is hard to justify from an economic point of view when using stochastic dynamic general equilibrium models. Along the following chapters, when

takling the issues related to the evaluation of the success of a model in replicating the stylized facts it wanted to explain, and to the comparison of the success of alternative model specifications, we have to bear in mind that a calibrator is not interested in verifying whether the model is "true" (in the sense of being the true DGP of the actual data) since the answer is already known from the outstart, but in identifying which aspects of the data a "false" model can replicate and whether different models give different answers because they are "false" in different dimensions.

Chapter 2 is a joint work with Fabio Canova, in which we illustrate the philosophy which forms the basis of calibration exercises in general equilibrium macroeconomic models and the details of the procedure, the advantages and the disadvantages of the approach, with particular reference to the issue of testing "false" economic models.

We provide an overview of the most recent simulation-based approaches to the testing problem and compare them to standard econometric methods used to test the fit of non-linear dynamic general equilibrium models. We illustrate how simulation-based techniques can be used to formally evaluate the fit of a calibrated model to the data and obtain ideas on how to improve the model design using a standard problem in the international real business cycle literature, i.e. whether a model with complete financial markets and no restrictions to capital mobility is able to reproduce the second order properties of aggregate saving and aggregate investment in an open economy (see Baxter and Crucini (1993)). The simulation-based methods suggested to assess the fit of a model (and to the related issue of solving the problem of giving values to the deep paremeters of the model) could prove useful when applied to any model although they are particularly built for testing calibrated models.

Chapter 3 proposes a new methodology to assess the fit of multivariate dynamic models whose solution is not exact but approximated. This is the case of most calibrated models but not only, so that it could be applied to any multivariate dynamic models. The approach is based on multivariate frequency domain techniques, which makes it especially suitable for models that focus on a particular frequency range, such as business cycle models.

An asymptotic test is presented for assessing how well a simulated model reproduces the dynamic properties of a vector of actual series. A further test is then derived

to compare the relative performance of alternative model specifications with respect to the multivariate vector of interest. The test is able to address the issue of comparing misspecified models, testing whether they are similar to each other while being different from the actual DGP. Monte Carlo evidence is provided to show the finite sample behavior of the tests, as well as their sensitivity to the sample size and the parametric structure of the DGP. Both tests are found to have high power even in small sample sizes. We also find, in common with other studies, that the small sample properties of our tests depend on the small sample characteristics of the spectral estimators employed.

The use of the methodology proposed is illustrated by evaluating to which extent different versions of a two-country two-good International Real Business Cycle based on Backus, Kehoe and Kydland (1993) (which differ on the degree of final goods trade, common shocks and spillovers across national disturbances) can reproduce the interdependencies observed between the US and European real GDPs at business cycle frequencies. Our model evaluation procedure confirms statistically the rejection of all models (suggested by informal comparison of the spectral properties of actual and model data) and manages to produce a clear ranking of competing models according to their fit, which could not be done in our case with simple inspection of spectral densities of the actual and simulated data.

Each of the model evaluation methodologies studied throughout Chapters 2 and 3 has been proposed as an alternative to the informal assessment of stochastic dynamic general equilibrium models. However, it is a difficult task for the researcher to choose which one to use. Their diversity makes it potentially possible that alternative methodologies assess as very different the success of a model in reproducing the same stylized facts of the actual data. A comparison under uniform conditions of the performance of these recently proposed methodologies is called for.

Chapter 4 "tests" the performance of the approaches of Watson (1993), DeJong, Ingram and Whiteman (1996), Canova and De Nicol'o (1995) and the spectral density distance approach suggested in Chapter 3 using Monte Carlo techniques. We ask: Do the different evaluation methodologies effectively improve the informal evaluation of a model by a "naive" calibrator? Are they only valid under limited assumptions, for

evaluating the fit over a particular set of statistics or a particular model? Our Monte Carlo experiments evaluate the ability of each methodology to accept a model when it is equal to the actual DGP and to reject it when it is at odds with the actual DGP. In a sense, we are treating each methodology as a test for stochastic dynamic general equilibrium models and compute its "size" and "power", respectively. We find that all four methodologies outperfrom the naive calibrator's rule since they substancially reduce the risk of rejecting the true DGP and are able to discriminate more clearly between the DGP and models different to it.

Chapter 2

Testing Calibrated General Equilibrium Models

2.1 Introduction

Simulation techniques are now used in many fields of applied research. They have been employed to compute estimators in situations where standard methods are impractical or fail, to evaluate the properties of parametric and nonparametric econometric estimators, to provide a cheap way of evaluating posterior integrals in Bayesian analysis and to undertake linear and nonlinear filtering with a computationally simple approach.

The task of this chapter is to describe how simulation based methods can be used to evaluate the fit of dynamic general equilibrium models specified using a calibration methodology, to compare and contrast their usefulness relative to more standard econometric approaches and to provide an explicit example where the various features of the approach can be highlighted and discussed.

The structure of this chapter is as follows. In Section 2.2 we provide a definition of what we mean by calibrating a model and discuss the philosophy underlying the approach and how it differs from standard dynamic time series modelling. We also discuss various approaches to the selection of model parameters, how to choose the vector of statistics used to compare actual with simulated data and how simulations are performed. Section 2.3 describes how to formally evaluate the model's approximation to the data and discusses alternative approaches to account for the uncertainty faced by

a simulator in generating time paths for the relevant variables. Although we present a general overview of alternative evaluation techniques, the focus is on simulation based methods. Section 2.4 briefly discusses how calibrated models can be used for policy analyses. In Section 2.5 we present an example, borrowed from Baxter and Crucini (1993), where the features of the various approaches to evaluation can be examined. Section 2.6 concludes.

2.2 What is Calibration?

2.2.1 A Definition

Although it is more than a decade since calibration techniques emerged in the main stream of dynamic macroeconomics (see Kydland and Prescott (1982)), a precise statement of what it means to calibrate a model has yet to appear in the literature. In general, it is common to think of calibration as an unorthodox procedure to select the parameters of a model. This need not to be the case since it is possible to view parameter calibration as a particular econometric technique where the parameters of the model are estimated using an "economic" instead of a "statistical" criteria (see e.g. Canova (1994)). On the other hand, one may want to calibrate a model because there is no available data to estimate its parameters, for example, if one is interested in studying the effect of certain taxes in a newly born country.

Alternatively, it is possible to view calibration as a cheap way to evaluate models. For example, calibration is considered by some a more formal version of the standard back-of-the-envelope calculations that theorists perform to judge the validity of their models (see e.g. Pesaran and Smith (1992)). According to others, calibration is a way to conduct quantitative experiments using models which are known to be "false", i.e. improper or simplified approximations of the true data generating processes of the actual data (see e.g. Kydland and Prescott (1991)).

Pagan (1994) stresses that the unique feature of calibration exercises does not lie so much in the way parameters are estimated, as the literature has provided alternative ways of doing so, but in the particular collection of procedures used to test tightly specified (and false) theoretical models against particular empirical facts. Here we

take a more general point of view and identify 6 steps which we believe capture the essence of the methodology. We call calibration a procedure which involves:

(i) Formulating an economic question to be addressed.

(ii) Selecting a model design which bears some relevance to the question asked.

(iii) Choosing functional forms for the primitives of the model and finding a solution for the endogenous variables in terms of the exogenous variables and the parameters.

(iv) Choosing parameters and stochastic processes for the exogenous variables and simulating paths for the endogenous variables of the model.

(v) Selecting a metric and comparing the outcomes of the model relative to a set of "stylized facts".

(vi) Doing policy analyses if required.

By "stylized facts" the literature typically means a collection of sample statistics of the actual data such as means, variances, correlations, etc., which (a) do not involve estimation of parameters and (b) are self-evident. More recently, however, the first requirement has been waived and the parameters of a VAR (or the impulse responses) have also been taken as the relevant stylized facts to be matched by the model (see e.g. Smith (1993), Cogley and Nason (1994)).

The next two subsections describe in details both the philosophy behind the first four steps and the practicalities connected with their implementation.

2.2.2 Formulating a question and choosing a model

The first two steps of a calibration procedure, to formulate a question of interest and a model which bears relevance to the question, are self evident and require little discussion. In general, the questions posed display four types of structure (see e.g. Kollintzas (1992) and Kydland (1992)):

- Is it possible to generate Z using theory W?

- How much of the fact X can be explained with impulses of type Y?
- What happens to the endogenous variables of the model if the stochastic process for the control variable V is modified ?
- Is it possible to reduce the discrepancy D of the theory from the data by introducing feature F in the model?

Two economic questions which have received considerable attention in the literature in the last 10 years are the so-called equity premium puzzle, i.e. the inability of a general equilibrium model with complete financial markets to quantitatively replicate the excess returns of equities over bonds over the last hundred years (see e.g. Mehra and Prescott (1985)) and how much of the variability of GNP can be explained by a model whose only source of dynamics are technology disturbances (see e.g. Kydland and Prescott (1982)). As is clear from these two examples, the type of questions posed are very specific and the emphasis is on the numerical implications of the exercise. Generic questions with no numerical quantification are not usually studied in this literature.

For the second step, the choice of an economic model, there are essentially no rules except that it has to have some relevance with the question asked. For example, if one is interested in the equity premium puzzle, one can choose a model which is very simply specified on the international and the government side, but very well specified on the financial side so that it is possible to calculate the returns on various assets. Typically, one chooses dynamic general equilibrium models. However, several authors have used model designs coming from different paradigms (see e.g. the neo-keynesian model of Gali (1994), the non-walrasian models of Danthine and Donaldson (1992) or Gali (1995) and the model with union bargaining of Eberwin and Kollintzas (1995)). There is nothing in the procedure that restricts the class of model design to be used. The only requirement is that the question that the researcher formulates is quantifiable within the context of the model and that the theory, in the form of a model design, is fully specified.

It is important to stress that a model is chosen on the basis of the question asked and not on its being realistic or being able to best replicate the data (see Kydland and Prescott (1991) or Kydland (1992)). In other words, how well it captures reality is not

a criteria to select models. What is important is not whether a model is realistic or not but whether it is able to provide a quantitative answer to the specific question the researcher poses.

This brings us to discuss an important philosophical aspect of the methodology. From the point of view of a calibrator all models are approximations to the DGP of the data and, as such, false. This aspect of the problem has been appreciated by several authors even before the appearance of the seminal article of Kydland and Prescott. For example, Hansen and Sargent (1979) also concede that an economic model is a false DGP for the data. Because of this and in order to test the validity of the model using standard statistical tools, they complete the probabilistic structure of the model by adding additional sources of variability, in the form of measurement errors or unobservable variables, to the fundamental forces of the economy.

For calibrators, the model is not a null hypothesis to be tested but an approximation of a few dimensions of the data. A calibrator is not interested in verifying whether the model is true (the answer is already known from the outstart), but in identifying which aspects of the data a false model can replicate and whether different models give different answers because they are false in different dimensions. A calibrator is satisfied with his effort if, through a process of theoretical respecification, a simple and highly stylized model captures an increasing number of features of the data (confront this activity with the so-called normal science of Kuhn (1970)).

Being more explicit, consider the realization of a vector of stochastic processes y_t (our data) and some well specified theoretical model $x_t = f(z_t, \gamma)$ which has something to say about y_t, where z_t are exogenous and predetermined variables and γ is a parameter vector. Because the model does not provide a complete description of the phenomenon under investigation we write

$$y_t = x_t + u_t \tag{2.1}$$

where u_t is an error representing what is missing from $f(z_t, \gamma)$ to reproduce the stochastic process generating y_t and whose properties are, in general, unknown (it need not necessarily be mean zero, serially uncorrelated, uncorrelated with the x's and so on). Let B_y and B_x be continuous and differentiable functions of actual and simulated data, respectively. Then standard econometric procedures judge the coherence of the model

to the data by testing whether or not $B_x = B_y$, given that the difference between B_x and B_y and their estimated counterpart $\hat{B}_x$ and $\hat{B}_y$ arise entirely from sampling error. While this is a sensible procedure when the null hypothesis is expected to represent the data, it is less sensible when it is known that the model does not completely capture all aspects of the data.

The third step of a calibration exercise concerns the solution of the model. To be able to obtain quantitative answers from a model it is necessary to find an explicit solution for the endogenous variables of the model in terms of the exogenous and predetermined variables and the parameters. For this reason it is typical to parameterize the objective function of the agents so that manipulation of the first order conditions is analytically tractable. For example, in general equilibrium models, it is typical to choose Cobb-Douglas production functions and constant relative risk aversion utility functions. However, although the main objective is to select simple enough functional forms, it is well known that almost all general equilibrium models and many partial equilibrium models have exact analytical solutions only in very special situations.

For general equilibrium models, a solution exists if the objective function is quadratic and the constraints linear (see e.g. Hansen and Sargent (1979)) or when the objective function is log-linear and the constraints linear (see e.g. Sargent (1987, ch.2)). In the other cases, analytical expressions relating the endogenous variables of the model to the "states" of the problem does not exist and it is necessary to resort to numerical techniques to find solutions which approximate equilibrium functionals either locally or globally. There has been substantial theoretical development in this area in the last few years and several solution algorithms have appeared in the literature (see e.g. the special January 1990 issue of the JBES or Marcet (1994)).

The essence of the approximation process is very simple. The exact solution of a model is a relationship between the endogenous variables x_t, the exogenous and predetermined variables z_t and a set of "deep" parameters γ of the type $x_t = f(z_t, \gamma)$ where f is generally unknown. The approximation procedures generate a relationship of the type $x_t^* = g(z_t, \gamma)$ and where $||f - g|| < \epsilon$ is minimal for some local or global metric. Examples of these types of procedures appear in Kydland and Prescott (1982), Coleman (1989), Tauchen and Hussey (1991), Novales (1990), Baxter (1992) and Marcet (1992),

among others. The choice of a particular approximation procedure depends on the question asked. If one is concerned with the dynamics of the model around the steady state, local approximations are sufficient. On the other hand, if one is interested in comparing economic policies requiring drastic changes in the parameters of the control variables, global approximation methods must be preferred.

2.2.3 Selecting Parameters and Exogenous Processes

Once an approximate solution has been obtained, a calibrator needs to select the parameters γ and the exogenous stochastic process z_t to be fed into the model in order to generate time series for x_t^*. There are several approaches to the choice of these two features of the model. Consider first the question of selecting z_t. This choice is relatively uncontroversial. One either chooses it on the basis of tractability or to give the model some realistic connotation. For example, one can assume that z_t is an AR process with innovations which are transformations of a $N(0,1)$ process and draw one or more realizations for z_t using standard random number generators. Alternatively, one can select the Solow residuals of the actual economy, the actual path of government expenditure or of the money supply. Obviously, the second alternative is typically preferred if policy analyses are undertaken. Note that while in both cases z_t is the realization of a stochastic process, in the first case the DGP is known while in the second it is not and this has implications for the way one measures the uncertainty in the outcomes of the model.

Next, consider the selection of the vector of parameters γ. Typically, they are chosen so that the model reproduces certain observations. Taking an example from physics, if one is interested in measuring water temperature in various situations it will be necessary to calibrate a thermometer for the experiments. For this purpose a researcher arbitrarily assigns the value 0 C to freezing water and the value 100 C to boiling water and interpolates values in the middle with, say, a linear scale. Given this calibration of the thermometer, one can then proceed to measure the results of the experiments: a value close to 100 C indicates "hot" water, a value close to 30 C indicates "tepid" water, and so on. To try to give answers to the economic question he poses, a calibrator must similarly select observations to be used to calibrate the model-

thermometer. There are at least three approaches in the literature. One can follow the deterministic computable general equilibrium (CGE) tradition, summarized, e.g. in Showen and Walley (1984), the dynamic general equilibrium tradition pioneered by Kydland and Prescott (1982) or employ more standard econometric techniques. There are differences between the first two approaches. The first one was developed for deterministic models which do not necessarily possess a steady state while the second one has been applied to dynamic stochastic models whose steady state is unique. Kim and Pagan (1994) provide a detailed analysis of the differences between these two approaches. Gregory and Smith (1993) supplement the discussion by adding interesting insights in the comparison of the first two approaches with the third.

In CGE models a researcher solves the model linearizing the system of equations by determining the endogenous variables around a hypothetical equilibrium where prices and quantities are such that there is no excess demand or excess supply. It is not necessary that this equilibrium exists. However, because the coefficients of the linear equations determining endogenous variables are functions of these equilibrium values, it is necessary to measure this hypothetical equilibrium. The main problem for this literature is therefore to find a set of "benchmark data" and to calibrate the model so that it can reproduce this data. Finding this data set is the most complicated part of the approach since it requires a lot of judgement and ingenuity. The process of specification of this data set leaves some of the parameters of the model typically undetermined, for example, those that describe the utility function of agents. In this situation a researcher either assigns arbitrary values or fixes them to values estimated in other studies in the literature. Although these choices are arbitrary, the procedure is coherent with the philosophy of the models: a researcher is interested in examining deviations of the model from a hypothetical equilibrium, not from an actual economy.

In stochastic general equilibrium models, the model is typically calibrated at the steady state: parameters are chosen so that the model, in the steady state, produces values for the endogenous variables which match corresponding long run averages of the actual data. In both this approach and the CGE approach point estimates of the parameters used to calibrate the model to the equilibrium are taken to be exact (no standard deviations are typically attached to these estimates). As in the previous

setup, the steady state does not necessarily pin down all the parameters of the model. Canova (1994) and Gregory and Smith (1993) discuss various methods to select the remaining parameters. Briefly, a researcher can choose parameters a-priori, pin them down using values previously estimated in the literature, can informally estimate them using simple method of moment conditions or formally estimate them using procedures like GMM (see e.g. Christiano and Eichenbaum (1992)), SMM (see e.g. Duffie and Singleton (1993)) or maximum likelihood (see e.g. McGratten, Rogerson and Wright (1993)). As pointed out by Kydland and Prescott (1991), choosing parameters using the information contained in other studies imposes a coherence criteria among various branches of the profession. For example, in the business cycle literature one uses stochastic growth models to examine business cycle fluctuations and checks the implications of the model using parameters typically obtained in micro studies, which do not employ data having to do with aggregate business cycle fluctuations (e.g. micro studies of labor markets).

If one follows a standard econometric approach, all the parameters are chosen by minimizing the MSE of the error u_t in (2.1), arbitrarily assuming that the error and the model designs are orthogonal, or by minimizing the distance between moments of the actual data and the model or maximizing the likelihood function of the data given the model design. As we already pointed out, this last approach is the least appealing one from the point of view of a calibrator since it makes assumptions on the time series properties of u_t which are hard to justify from an economic point of view.

To clearly understand the merits of each of these procedures it is useful to discuss their advantages and their disadvantages. Both the CGE and the Kydland and Prescott approach where some of the parameters are chosen a-priori or obtained from a very select group of studies are problematic in several respects. First, there is a selectivity bias problem (see Canova (1995)). There exists a great variety of estimates of the parameters in the literature and different researchers may refer to different studies even when they are examining the same problem. Second, there is a statistical inconsistency problem which may generate very spurious and distorted inference. As Gregory and Smith (1989) have shown, if some parameters are set a-priori and others estimated by simulation, estimates of the latter may be biased and inconsistent unless the parameters

of the former group are the true parameters of the DGP or consistent estimates of them. Third, since any particular choice is arbitrary, extensive sensitivity analysis is necessary to evaluate the quality of the results. To solve these problems Canova (1994)-(1995) suggests an approach for choosing parameters which allows, at a second stage, to draw inferences about the quality of the approximation of the model to the data. The idea is very simple. Instead of choosing one set of parameters over another he suggests calibrating each parameter of the model to an interval, using the empirical information to construct a distribution over this interval (the likelihood of a parameter given existing estimates) and conducting simulation by drawing parameter vectors from the corresponding joint "empirical" distribution. An example may clarify the approach. If one of the parameters of interest is the coefficient of constant relative risk aversion of the representative agent, one typically chooses a value of 2 and tries a few values above and below this one to see if results change. Canova suggests taking a range of values, possibly dictated by economic theory, say [0,20], and then over this range constructing a histogram using existing estimates of this parameter. Most of the estimates are in the range [1,2] and in some asset pricing models researchers have tried values up to 10. Given this information, the resulting empirical distribution for this parameter can be very closely approximated by a $\chi^2(2)$, which has the mode at 2 and about 5% probability in the region above 6.

The selection of the parameters of theoretical models through statistical estimation has advantages and disadvantages. The main advantage is that these procedures avoid arbitrary choices and explicitly provide a measure of dispersion for the estimates which can be used at a second stage to evaluate the quality of the approximation of the model to the data. The disadvantages are of various kinds. First of all, to undertake a formal or informal estimation it is typically necessary to select the moments one wants to fit, and this choice is arbitrary. The standard approach suggested by Kydland and Prescott can indeed be thought of as a method of moment estimation where one chooses parameters so as to set only the discrepancy between the first moment of the model and the data (i.e. the long run averages) to zero. The formal approach suggested by Christiano and Eichenbaum (1992) or Langot and Fève (1994), on the other hand, can be thought of as a method of moment estimation where a researcher

fits the discrepancies between model and data first and second moments to zero. The approach of choosing parameters by setting to zero the discrepancy between certain moments has the disadvantage of reducing the number of moments over which it will be possible to evaluate the quality of the model. Moreover, it is known that estimates obtained with the method of moments or GMM may be biased. Therefore, simulations and inference conducted with these estimates may lead to spurious inference (see e.g. Canova, Finn and Pagan (1994)). In addition, informal SMM may lead one to select parameters even though they are not identifiable (see Gregory and Smith (1989)). Finally, one should note that the type of uncertainty which is imposed on the model via an estimation process does not necessarily reflect the uncertainty a calibrator faces when choosing the parameter vector. As is clear from a decade of GMM estimation, once the moments are selected and the data given, sample uncertainty is pretty small. The true uncertainty is in the choice of moments and in the data set to be used to select parameters. This uncertainty is disregarded when parameters are chosen using extremum estimators like GMM.

Finally, it is useful to compare the parameter selection process used by a calibrator à-la Kydland and Prescott and the one used by a traditional econometric approach. In a traditional econometric approach parameters are chosen so as to minimize some **statistical** criteria, for example, the MSE. Such criteria do not have any economic content, impose stringent requirements on the structure of u_t and are used, primarily, because there exists a well established statistical and mathematical literature on the subject. In other words, the parameter selection criteria used by traditional econometricians does not have economic justification. On the other hand, the parameter selection criteria used by followers of the Kydland and Prescott methodology can be thought of as being based on **economic** criteria. For example, if the model is calibrated so that, in the steady state, it matches the long run features of the actual economy, parameters are implicitly selected using the condition that the sum (over time) of the discrepancies between the model and the data is zero. In this sense there is an important difference between the two approaches which has to do with the assumptions that one is willing to make on the errors u_t. By calibrating the model to long run observations a researcher selects parameters assuming $E(u) = 0$, i.e. using a restriction which is identical to

the one imposed by a GMM econometrician who chooses parameters using only first moment conditions. On the other hand, to conduct classical inference a researcher imposes restrictions on the first and second moments of u_t.

The comparison we have made so far concerns, obviously, only those parameters which enter the steady state conditions of the model. For the other parameters a direct comparison with standard econometric practice is not possible. However, if all parameters are calibrated to intervals with distributions which are empirically determined, the calibration procedure we have described shares a tight connection with Bayesian inferential methods such as Consensus Analysis or Meta-Analysis (see e.g. Genest and Zidak (1986) or Wolf (1986)).

Once the parameters and the stochastic processes for the exogenous variables are selected and an (approximate) solution to the model has been found, simulated paths for x_t^* can be generated using standard Monte Carlo simulation techniques.

2.3 Evaluating Calibrated Models

The questions of how well a model matches the data and how much confidence a researcher ought to give to the results constitute the most crucial steps in the calibration procedure. In fact, the most active methodological branch of this literature concerns methods to evaluate the fit of a model selected according to the procedures described in Section 1.2. The evaluation of a model requires three steps: first, the selection of a set of stylized facts; second, the choice of a metric to compare functions of actual and simulated data and third, the (statistical) evaluation of the magnitude of the distance. Formally, let S_y be a set of statistics (stylized facts) of the actual data and let $S_{x^*}(z_t, \gamma)$ be a set of statistics of simulated data, given a vector of parameters γ and a vector of stochastic processes z_t. Then model evaluation consists of selecting a function $\psi(S_y, S_{x^*}(z_t, \gamma))$ measuring the distance between S_y and S_{x^*} and in assessing its magnitude.

The choice of which stylized facts one wants to match obviously depends on the question asked and on the type of model used. For example, if the question is what is the proportion of actual cyclical fluctuations in GNP and consumption explained by the model, one would choose stylized facts based on variances and covariances of the

data. As an alternative to the examination of second moments, one could summarize the properties of actual data via a VAR and study the properties of simulated data, for example, by comparing the number of unit roots in the two sets of data (as in Canova, Finn and Pagan (1994)), the size of VAR coefficients (as in Smith (1993)) or the magnitude of certain impulse responses (as in Cogley and Nason (1994)). Also, it is possible to evaluate the discrepancy of a model to the data by choosing specific events that one wants the model to replicate e.g. business cycle turning points, (as in King and Plosser (1994) or Simkins (1994)) or variance bounds (as in Hansen and Jagannathan (1991)).

Classical pieces in the calibration literature (see e.g. Kydland and Prescott (1982) or (1991)) are typically silent on the metric one should use to evaluate the quality of the approximation of the model to the data. The approach favored by most calibrators is to glare over the exact definition of the metric used and informally assess the properties of simulated data by comparing them to the set of stylized facts. In this way a researcher treats the computational experiment as a measurement exercise where the task is to gauge the proportion of some observed statistics reproduced by the theoretical model. This informal approach is also shared by cliometricians (see e.g. Summers (1991)) who believe that rough reproduction of simple sample statistics is all that is needed to evaluate the implications of the model ("either you see it with naked eyes or no fancy econometric procedure will find it").

There are, however, alternatives to this informal approach. To gain some understanding of the differences among approaches, but at the cost of oversimplifying the matter, we divide evaluation approaches into five classes:

- Informal approaches.

- Approaches which do not consider sampling variability of actual or the uncertainty in simulated data, but instead use the statistical properties of u_t in (2.1) to impose restrictions on the time series properties of ψ. This allows them to provide an R^2-type measure of fit between the model and the data (see Watson (1993)).

- Approaches which use the sampling variability of the **actual** data (affecting S_y

and, in some cases, estimated γ) to provide a measure of the distance between the model and the data. Among these we list the GMM based approach of Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1993) or Fève and Langot (1994), and the frequency domain approaches of Diebold, Ohanian and Berkowitz (1995) and that explained in Chapter 2 of this dissertation.

- Approaches which use the uncertainty of the **simulated** data to provide a measure of distance between the model and the data. Among these procedures we can distinguish those who take z_t as stochastic and γ as given, such as Gregory and Smith (1991), Söderlind (1994) or Cogley and Nason (1994) and those who take both z_t and γ as stochastic, such as Canova (1994) and (1995).

- Finally, approaches which consider the sampling variability of the **actual** data and the uncertainty in **simulated** data to evaluate the fit of the model. Once again we can distinguish approaches which, in addition to taking S_y as random, allow for variability in the parameters of the model (keeping z_t fixed) such as DeJong, Ingram and Whiteman (1996) from those which allow for both z_t and γ to vary such as Canova and De Nicoló (1995).

Because we want to put the emphasis of this chapter on simulation techniques, we will only briefly examine the first three approaches and discuss in more detail the last two, which make extensive use of simulation techniques to conduct inference. Kim and Pagan (1994) provide a thorough critical review of several of these evaluation techniques and additional insights on the relationship among them.

The evaluation criteria that each of these approaches proposes is tightly linked to the parameter selection procedure we discussed in the previous section.

As mentioned the standard approach is to choose parameters using steady state conditions. Those parameters which do not appear in the steady state are selected a-priori or with reference to existing literature. Also, since S_y is chosen to be a vector of numbers and no uncertainty is allowed in the selected parameter vector, one is forced to use an informal metric to compare the model to the data. This is because, apart from the uncertainty present in the exogenous variables, the model links the endogenous variables to the parameters in a deterministic fashion. Therefore, once

we have selected the parameters and we have a realization of S_y, it is not possible to measure the dispersion of the distance $\psi(S_y, S_{x^*}(z_t, \gamma))$. From the point of view of the majority of calibrators this is not a problem. As emphasized by Kydland and Prescott (1991) or Kydland (1992), the trust a researcher has in an answer given by the model does not depend on a statistical measure of discrepancy, but on how much he believes in the economic theory used and in the measurement undertaken.

Taking this as the starting point of the analysis Watson (1993) suggests an ingenious way to evaluate models which are known to be an incorrect DGP for the actual data. Watson asks how much error should be added to x_t^* so that its autocovariance function equals the autocovariance function of y_t. Writing $y_t = x_t^* + u_t^*$ where u_t^* includes both model error u_t and the approximation error due to the use x_t^* in place of x_t, the autocovariance function of this error is given by

$$A_{u^*}(z) = A_y(z) + A_{x^*}(z) - A_{x^*y}(z) - A_{yx^*}(z) \tag{2.2}$$

To evaluate the last two terms in (2.2) we need a sample from the joint distribution of (x_t^*, y_t) which is not available. In these circumstances it is typical to assume that either u_t^* is a measurement error or a signal extraction noise (see e.g. Sargent (1989)), but in the present context neither of the two assumptions is very appealing. Watson suggests choosing $A_{x^*y}(z)$ so as to minimize the variance of u_t^* subject to the constraint that $A_{x^*}(z)$ and $A_y(z)$ are positive semidefinite. Intuitively, the idea is to select $A_{x^*y}(z)$ to give the best possible fit between the model and the data (i.e. the smallest possible variance of u_t^*). The exact choice of $A_{x^*y}(z)$ depends on the properties of x_t^* and y_t, i.e. whether they are serially correlated or not, scalar or vectors, full rank processes or not. In all cases, the selection criteria chosen imply that x_t^* and y_t are perfectly linearly correlated where the matrix linking the two vectors depends on their time series properties and on the number of shocks buffeting the model. Given this framework of analysis, Watson suggests two measures of fit, similar to a $1 - R^2$ from a regression, of the form

$$r_j(\omega) = \frac{A_{u^*}(\omega)_{jj}}{A_y(\omega)_{jj}}, \quad \forall \omega \tag{2.3}$$

$$R_j = \frac{\int_{\omega \in Z} A_{u^*}(\omega)_{jj} d\omega}{\int_{\omega \in Z} A_y(\omega)_{jj} d\omega} \tag{2.4}$$

where the first statistic measures the variance of the j-th variable in the error vector relative to the variance of the j-th variable in the vector of actual data for each frequency and the second statistic is the sum of the first over a set of frequencies. This last measure may be useful to evaluate the model, say, at business cycle frequencies. It should be stressed that (2.3) and (2.4) are lower bounds. That is, when $r_j(\omega)$ or $R_j(\omega)$ are large, the model poorly fits the data. However, when they are small, it does not necessarily follow that the model is appropriate since it may still fit the data poorly if we change the assumptions about $A_{x\bullet y}(z)$.

To summarize, Watson chooses the autocovariance function of y as the set of stylized facts of the data to be matched by the model, the ψ function as the ratio of $A_{u\bullet}$ to A_y and evaluates the size of ψ informally (i.e. if it is greater than one, between zero and one or close to zero). Note that in this approach, γ and z_t are fixed, and $A_{x\bullet}$ and A_y are assumed to be measured without error.

When a calibrator is willing to assume that parameters are measured with error because, given an econometric technique and a sample, parameters are imprecisely estimated, then model evaluation can be conducted using measures of dispersion for simulated statistics which reflect parameter uncertainty. There are various versions of this approach. Christiano and Eichenbaum (1992), Cecchetti, Lam and Mark (1993) and Fève and Langot (1994) use a version of a J-test to evaluate the fit of a model. In this case S_y are moments of the data while ψ is a quadratic function of the type

$$\psi(S_y, S_{x\bullet}(z_t, \gamma)) = [S_y - S_{x\bullet}(\gamma)]V^{-1}[S_y - S_{x\bullet}(\gamma)]' \tag{2.5}$$

where V is a matrix which linearly weights the covariance matrix of $S_{x\bullet}$ and S_y, and $S_{x\bullet}$ is random because γ is random. Formal evaluation of this distance can be undertaken following Hansen (1982): under the null that $S_y = S_{x\bullet}(z_t, \gamma)$ the statistic defined in (2.5) is asymptotically distributed as a χ^2 with the number of degrees of freedom equal to the number of overidentifying restrictions, i.e. the dimension of S_y minus the dimension of the vector γ. Note that this procedure is correct asymptotically, that it implicitly assumes that $x_t = f(z_t, \gamma)$ (or its approximation x_t^*) is the correct DGP for the data and that the relevant loss function measuring the distance between actual and simulated data is quadratic.

The methods proposed by Diebold, Ohanian and Berkowitz (DOB) (1994) and the one presented in Chapter 2 are slightly different but can be broadly included into this class of approaches.

For DOB the statistic of interest is the spectral density matrix of y_t and, given a sample, this is assumed to be measured with error. They measure the uncertainty surrounding point estimates of the spectral density matrix employing (small sample) 90% confidence bands constructed using parametric and nonparametric bootstrap approaches and Bonferroni tunnels. On the other hand, they take the realization of z_t as given so that the spectral density matrix of simulated data can be estimated without error simply by simulating very long time series for x_t^*, and they estimate the parameters of the model so that they generate the best fit, i.e. so that they minimize the distance between model and actual data spectral density matrices. The approach presented in Chapter 2 also takes the spectral density matrix as the set of stylized facts of the data to be matched by the model. Unlike DOB, it considers the uncertainty in actual and simulated data by jointly estimating the spectral density matrix of actual and simulated data and constructs measures of uncertainty around point estimates of the spectral density matrix using asymptotic distribution theory.

In both cases, the measure of fit used is generically given by:

$$C(\gamma, z_t) = \int_0^\pi \psi(F_y(\omega), F_{x^*}(\omega, \gamma, z_t))W(\omega)d\omega \tag{2.6}$$

where $W(\omega)$ is a set of weights applied to different frequencies and F are the spectral density matrices of actual and simulated data.

DOB suggest various options for ψ (quadratic, ratio, likelihood type) but do not construct a direct test statistic to examine the magnitude of ψ. Instead, they compute a small sample distribution of the event that $C(\gamma, z_t)$ is close to a particular value (zero if ψ is quadratic, 1 if ψ is a ratio, etc.) The approach in Chapter 2, on the other hand, explicitly uses a quadratic expression for ψ and uses an asymptotic χ^2 test to assess whether the magnitude of the discrepancy between the model and the data is significant or not. The set of asymptotic tools developped can also be used to compare the fit of two alternative models to the data and decide which one is more acceptable.

If a calibrator is willing to accept the idea that the stochastic process for the

exogenous variables is not fixed, she can then compute measures of dispersion for simulated statistics by simply changing the realization of z_t while maintaining the parameters fixed. Such a methodology has its cornerstone in the fact that it is the uncertainty in the realization of the exogenous stochastic process (e.g. the technology shock), an uncertainty which one can call extrinsic, and not the uncertainty in the parameters, which one can call intrinsic, which determines possible variations in the statistics of simulated data. Once a measure of dispersion of simulated statistics is obtained, the sampling variability of simulated data can be used to evaluate the distance between statistics of actual and simulated data (as e.g. Gregory and Smith (1991) and (1993)).

If one uses such an approach, model evaluation can be undertaken with a probabilistic metric using well known Monte Carlo techniques. For example, one may be interested in finding out in what decile of the simulated distribution the actual value of a particular statistic lies, in practice, calculating the "size" of calibration tests. This approach requires two important assumptions: that the evaluator takes the model economy as the true DGP for the data and that differences between S_y and $S_{x\bullet}$ occur only because of sampling variability. To be specific, Gregory and Smith take S_y be a set of moments of the data and assume that they can be measured without error. Then, they construct a distribution of $S_{x\bullet}(z_t, \gamma)$ by drawing realizations for the z_t process from a given distribution, given γ. The metric ψ used is probabilistic, i.e. they calculate the probability $Q = P(S_{x\bullet} \leq S_y)$, and judge the fit of the model informally, e.g. measuring how close Q is to 0.5.

An interesting variation on this setup is provided by Söderlind (1994) and Cogley and Nason (1994). Söderlind employs the spectral density matrix of the actual data while Cogley and Nason choose a "structural" impulse response function as the relevant statistics to be matched. Söderlind maintains a probabilistic metric and constructs the empirical rejection rate for the event that the actual spectral density matrix of y_t lies inside the asymptotic 90% confidence band for the spectral density matrix of the simulated data. Such an event is replicated by drawing vectors z_t for a given distribution. Cogley and Nason choose a quadratic measure of distance which, under the null that the model is the DGP for the data, has an asymptotic χ^2 distribution

and then tabulate the empirical rejection rates of the test, by repeatedly constructing the statistic drawing realizations of the z_t vector. To be specific, the ψ function is in this case given by

$$\psi_{k,j}(\gamma) = [IRF^k_{x^*}(z^j_t, \gamma) - IRF^k_y]V^{-1}[IRF^k_{x^*}(z^j_t, \gamma) - IRF^k_y]' \tag{2.7}$$

where j indexes replications and k steps, IRF^k is the impulse response function and V is its asymptotic covariance matrix at step k. Because for every k and for fixed j $\psi_{k,j}(\gamma)$ is asymptotically χ^2, they can construct (a) the simulated distribution for $\psi_{k,j}$ and compare it with a χ^2 and (b) the rejection frequency for each model specification they examine.

In practice, all three approaches are computer intensive and rely on Monte Carlo methods to conduct inference. Also, it should be stressed that all three methods verify the validity of the model by computing the "size" of the calibration tests, i.e. assuming that the model is the correct DGP for y_t.

The approach of Canova (1994)-(1995) also belongs to this category of methods, but, in addition to allowing the realization of the stochastic process for the exogenous variables to vary, he also allows for parameter variability in measuring the dispersion of simulated statistics. The starting point, as discussed earlier, is that parameters are uncertain not so much because of sample variability, but because there are many estimates of the same parameter obtained in the literature, since estimation techniques, samples and frequency of the data tend to differ. If one calibrates the parameter vector to an interval, rather than to a particular value, and draws values for the parameters from the empirical distribution of parameter estimates, it is then possible to use the intrinsic uncertainty, in addition to or instead of the extrinsic one, to evaluate the fit of the model. The evaluation approach used is very similar to the one of Gregory and Smith: one simulates the model repeatedly by drawing parameter vectors from the empirical "prior" distribution and realizations of the exogenous stochastic process z_t from some given distribution. Once the empirical distribution of the statistics of interest is constructed, one can then compute either the size of calibration tests or the percentiles where the actual statistics lie.

The last set of approaches considers the uncertainty present in the statistics of both actual and simulated data to measure the fit of the model to the data. In essence what these approaches attempt to formally measure is the degree of overlap between the (possibly) multivariate distributions of S_y and S_x using Monte Carlo techniques. There are differences in the way these distributions have been constructed in the literature. Canova and De Nicoló (1995) use a parametric bootstrap algorithm to construct distributions for the statistics of the actual data . DeJong, Ingram and Whiteman (DIW) (1996), on the other hand, suggest representing the actual data with a VAR and computing posterior distribution estimates for the moments of interest by drawing VAR parameters from their posterior distribution and using the AR(1) companion matrix of the VAR at each replication. In constructing distributions of simulated statistics, Canova and De Nicoló take into account both the uncertainty in exogenous processes and parameters while DIW only consider parameter uncertainty. The two approaches also differ in the way the "prior" uncertainty in the parameters is introduced in the model. The former paper follows Canova (1995) and chooses empirical based distributions for the parameter vector. DIW use subjectively specified prior distributions (generally normal) whose location parameter is set at the value typically calibrated in the literature while the dispersion parameter is free. The authors use this parameter in order to (informally) minimize the distance between actual and simulated distributions of the statistics of interest. By enabling the specification of a sequence of increasingly diffuse priors over the parameter vector, such a procedure illustrates whether the uncertainty in the model's parameters can mitigate differences between the model and the data.

Finally, there are differences in assessing the degree of overlap of the two distributions. Canova and De Nicoló choose a particular contour probability for one of the two distributions and ask how much of the other distribution is inside the contour. In other words, the fit of the model is examined very much in the style of the Monte Carlo literature: a good fit is indicated by a high probability covering of the two regions. To describe the features of the two distributions, they also repeat the exercise varying the chosen contour probability, say, from 50% to 75%, 90%, 95% and 99%. The procedure allows them to detect anomalies in the shape of the two distributions due to clustering

of observations in one area, skewness or leptokurtic behavior. In this approach actual data and simulated data are used symmetrically in the sense that one can either ask whether the actual data could be generated by the model, or viceversa, whether simulated data are consistent with the distribution of the empirical sample. This symmetry allows the researcher to understand much better the distributional properties of error u_t in (2.1). Moreover, the symmetry with which the two distributions are treated resembles very much the process of switching the null and the alternative in standard classical hypothesis testing.

DeJong, Ingram and Whiteman take the point of view that there are no well established criteria to judge the adequacy of a model's "approximation" to reality. For this reason they present two statistics aimed at synthetically measuring the degree of overlap among distributions. One, which they call Confidence Interval Criterion (CIC) is the univariate version of the contour probability criteria used by Canova and De Nicoló and is defined as

$$CIC_{ij} = \frac{1}{1-\alpha}\int_a^b P_j(s_i)ds_i \tag{2.8}$$

where $s_i, \quad i = 1,\ldots,n$ is a set of functions of interest, $a = \frac{\alpha}{2}$ and $b = 1 - a$ are the quantiles of $D(s_i)$, the distribution of the statistic in the actual data, $P_j(s_i)$ is the distribution of the simulated statistic where j is the diffusion index of the prior on the parameter vector and $1 - \alpha = \int_a^b D(s_i)ds_i$. Note that with this definition, CIC_{ij} ranges between 0 and $\frac{1}{1-\alpha}$. For CIC close to zero, the fit of the model is poor, either because the overlap is small or because P_j is very diffuse. For CIC close to $\frac{1}{1-\alpha}$ the two distributions overlap substantially. Were the two distributions equal, CIC would be 1. If $CIC > 1$, $D(s_i)$ is diffuse relative to $P_j(s_i)$, i.e. the data is found to be relatively uninformative regarding s_i.

DeJong, Ingram and Whiteman suggest a second summary measure analogous to a t-statistic for the mean of $P_j(s_i)$ in the $D(s_i)$ distribution which complements the CIC measure, i.e.,

$$d_{ji} = \frac{EP_j(s_i) - ED(s_i)}{\sqrt{varD(s_i)}} \tag{2.9}$$

Large values of (2.9) indicate that the location of $P_j(s_i)$ is quite different from the location of $D(s_i)$. This difference of means measure allows to distinguish among the two

possible interpretations when CIC is close to zero, or to recognise model distributions $P_j(s_i)$ different to actual data ones (skewed or leptokurtic with respect to $D(s_i)$) even when the percentage ovelap is high.

The final problem of the DIW methodology is to choose α. DeJong, Ingram and Whiteman fix a particular value ($\alpha = 0.01$) but, as in Canova and De Nicoló, varying α for a given j is probably a good thing to do in order to describe the feature of the distributions. This is particularly useful when we are interested in partitions of the joint distributions of s_i because graphical methods or simple statistics are not particularly informative about distributions in high dimensional spaces.

2.4 Policy Analyses

Although it is not the purpose of this chapter to discuss in detail how calibrated models can be used for policy analyses, it is useful to describe the implications of the procedure for questions which have policy implications and how policy experiments can be undertaken. As we have already mentioned, a model is typically calibrated to provide a quantitative answer to very precise questions and some of these questions have potential policy implications. To forcefully argue the policy implications of the exercise one needs to be confident in the answer given by the model and to do this it is necessary to undertake extensive sensitivity analysis to check how results change when certain assumptions are modified.

As we have seen, the answers of the model come in the form of continuous functions $h(x_t^*) = h(g(z_t, \gamma))$ of simulated data. In theory, once g has been selected, the uncertainty in h is due to the uncertainty in γ and in z_t. Since in standard calibration exercises the γ vector is fixed, it is therefore typical to examine the sensitivity of the results in the neighborhood of the calibrated values for γ. Such experiments may be local, if the neighborhood is small, or global, in which case one measures the sensitivity of the results to perturbations of the parameters over the entire range. This type of exercise may provide two types of information. First, if results are robust to variations of a parameter in a particular range, its exact measurement is not crucial. In other words, the uncertainty present in the choice of such a parameter does not make the answers of the model tenuous and economic inference groundless. On the other hand,

if results crucially depend on the exact selection of certain parameters, it is clearly necessary to improve upon existing measurement of these parameters.

A local sensitivity analysis can be undertaken informally, replicating the experiments for different values of the parameters (as in Kydland and Prescott (1982)) or more formally, calculating the elasticity of h with respect to γ (as in Pagan and Shannon (1985)). A global sensitivity analysis can be efficiently undertaken with Monte Carlo methods or numerical semi-deterministic techniques (see e.g. Niederreiter (1988)) if the function g is known and the distribution of the γ vector is specified. If g is only an approximation to the functional linking x to z and γ, one can use techniques like Importance Sampling (see Geweke (1989)) to take into account this additional source of uncertainty. Clearly the two types of sensitivity analysis are not incompatible and should both be undertaken to assess the degree of trust a researcher can attach to the answer given by the model. Finally, one should note that the type of sensitivity analysis one may want to undertake depends also on the way parameters are selected and models evaluated. For example, if one uses the approach of Canova (1994)-(1995) or DeJong, Ingram and Whiteman (1995), the evaluation procedure automatically and efficiently provides sensitivity analysis to global perturbations of the parameters within an economically reasonable range.

Once model answers to the question of interest have been shown to be robust to reasonable variations in the parameters, a researcher may undertake policy analyses by changing the realization of the stochastic process for z_t or varying a subset of the γ vector, which may be under the control of, say, the government. Analyses involving changes in the distribution of z_t in the g function are also possible, but care should be exercised in order to compare results across specifications.

2.5 An example

In the field of international economics, robust stylized facts are usually hard to obtain. One of the most stable regularities observed in the data is the high correlation of national saving and domestic investment, both in time series analysis of individual countries and in cross sections regressions where the average over time of these variables is treated as a single data point for each country. High saving and investment

correlations are observed in small economies as well as large ones, although the correlation tends to be lower for smaller countries. These findings were originally interpreted as indicating that the world economy is characterized by a low degree of capital mobility. Yet most economists believe that the world is evolving toward an increasingly higher degree of international capital mobility. Baxter and Crucini (1993) forcefully turned this initial interpretation around by providing a model in which there is perfect international mobility of financial and physical capital but which generates high time series correlations of national saving and investment. Their evaluation of the model lies entirely within the standard Kydland and Prescott approach, i.e. parameters are fixed at some reasonably chosen values, no uncertainty is allowed in actual and simulated statistics and the metric used to compare actual and simulated data is informal.

The task of this section is three fold. First, we want to study whether the time series properties of simulated saving and investment do indeed reproduce those of the actual data when the model is formally examined with the tools described in this article. To this end we provide several measures of fit which can be used to gauge the closeness of the model to the data using variants of the simulation-based procedures described in the previous section. Second, we wish to contrast the outcomes obtained with various evaluation procedures and compare them with those obtained using more standard techniques. This will shed further light on the degree of approximation of the model to the data, and point out, when they emerge, unusual features of the model. Finally, we wish to provide a few suggestions on how to fine tune the model design so that undesiderable features are eliminated while maintaining the basic bulk of the results.

2.5.1 The model

We consider a model with two countries and a single consumption good. Each country is populated by a large number of identical agents and labor is assumed to be immobile across countries and variables are measured in per-capita terms. Preferences of the representative agent of country $h = 1, 2$ are given by:

$$U \equiv E_0 \sum_{t=0}^{\infty} \frac{\beta^t}{1-\sigma} [C_{ht}^{\mu} L_{ht}^{(1-\mu)}]^{1-\sigma} \tag{2.10}$$

where C_{ht} is private consumption of the single composite good by the representative agent of country h and L_{ht} is leisure, β is the discount factor, σ the coefficient of relative risk aversion and μ the share of consumption in utility. Leisure choices are constrained by:

$$0 \leq L_{ht} + N_{ht} \leq 1 \qquad \forall\, h \tag{2.11}$$

where the total endowment of time in each country is normalized to 1 and N_t represents the number of hours worked. The goods are produced with a Cobb-Douglas technology:

$$Y_{ht} = A_{ht}(K_{ht})^{1-\alpha}(X_{ht}N_{ht})^{\alpha} \qquad h = 1,2 \tag{2.12}$$

where K_t is the capital input, α is the share of labor in GDP, and where $X_{ht} = \theta_x X_{ht-1}\ \forall h$ with $\theta_x \geq 1$. X_{ht} represents labor-augmenting Harrod-neutral technological progress with deterministic growth rate equal to θ_x. Production requires domestic labor and capital inputs and is subject to a technological disturbance A_{ht} with the following properties:

$$\begin{bmatrix} A_{1t} \\ A_{2t} \end{bmatrix} = \begin{bmatrix} \bar{A}_1 \\ \bar{A}_2 \end{bmatrix} + \begin{bmatrix} \rho & \nu \\ \nu & \rho \end{bmatrix} \begin{bmatrix} A_{1t-1} \\ A_{2t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

where $\epsilon_t = [\epsilon_{1t}\ \epsilon_{2t}]' \sim N(0, \begin{bmatrix} \sigma_\epsilon^2 & \psi \\ \psi & \sigma_\epsilon^2 \end{bmatrix})$ and $[\bar{A}_1, \bar{A}_2]'$ is a vector of constants. The parameter ψ controls the contemporaneous spillover while ν the lagged spillover of the shocks.

Capital goods are accumulated according to:

$$K_{ht+1} = (1-\delta_h)K_{ht} + \phi(I_{ht}/K_{ht})K_{ht} \qquad h = 1,2 \tag{2.13}$$

where $\phi(\frac{I_{ht}}{K_{ht}}) > 0$ is concave and represents the costs of adjusting capital. As explained in Baxter and Crucini (1993), there is no need to choose a functional form for ϕ; it is sufficient to describe its behavior near the steady state. We do this by specifying two parameters: $\frac{1}{\phi'}$, which corresponds to Tobin's Q, i.e. the price of existing capital in one location relative to the price of new capital and $\xi_{\phi'}$, the elasticity of the marginal adjustment cost function with respect to the investment-capital ratio.

Governments finance their consumption purchases, G_{ht}, by taxing national outputs with a distorting tax and transferring what remains back to domestic residents. For

simplicity we assume that $G_{ht} = G_h$, $\forall t$. The government budget constraint is given by:

$$G_h = TR_{ht} + \tau_h Y_{ht} \qquad \forall\, h \tag{2.14}$$

where τ_h are tax rates and TR_h are lump sum transfers in country h.

The economy wide resource constraint is given by:

$$\pi(Y_{1t} - G_{1t} - C_{1t} - I_{1t}) + (1-\pi)(Y_{2t} - G_{2t} - C_{2t} - I_{2t}) \geq 0 \tag{2.15}$$

where π is the fraction of world population living in country 1.

Finally, following Baxter and Crucini (1993), we assume complete financial markets and free mobility of financial capital across countries so that agents can write and trade every kind of contingent security.

To find a solution to the model we first detrend those variables which drift over time by taking ratios of the original variables with respect to the labor augmenting technological progress, e.g. $y_{ht} = \frac{Y_{ht}}{X_{ht}}$, etc. Second, since there are distortionary taxes in the model, the competitive equilibrium is not Pareto optimal and the competitive solution differs from the social planner's solution. As in Baxter and Crucini (1993) we solve the problem faced by a pseudo social planner, modifying the optimality conditions to take care of the distortions. The weights in the social planner problem are chosen to be proportional to the number of individuals living in each of the countries. The modified optimality conditions are approximated with a log-linear expansion around the steady state as in King, Plosser and Rebelo (1988). Time series for saving and investment in each of the two countries are computed analytically from the approximate optimality conditions. The second order properties of saving and investment of actual and simulated data are computed eliminating from the raw time series a linear trend.

The parameters of the model are $\gamma = [\beta,\ \sigma,\ \alpha,\ \theta_x,\ \delta,\ \rho,\ \nu,\ \sigma_\epsilon,\ \psi,\ \pi,\ \xi_{\phi'},\ \phi',\ \tau]$ plus steady state hours and the steady state Tobin's Q which we set equal to 1. The exogenous processes of the model are the two productivity disturbances so that $z_t = [A_{1t}\ A_{2t}]'$.

The actual data we use are per capita basic saving (i.e. computed as $S_t = Y_t - C_t - G_t$) and investment for the period 1970Q1–1993Q3 for the US and for Europe in real terms, seasonally adjusted and are from OECD Main Economic Indicators. Plots of the detrended series appear in Figure 1.1.

The statistics we care about are the diagonal elements of the 4×4 spectral density matrix of the data and the coherences between saving and investment of the two "countries". Spectral density estimates at each frequency are computed smoothing with a flat window 13 periodogram ordinates. Figure 1.2 plots these statistics.

In the benchmark experiment the vector γ is the same as in Baxter and Crucini (1993) except for σ_ϵ which they normalize to 1, while we set it equal to the value used in Backus, Kehoe and Kydland (1995), and are reported in the first column of Table 1.1. When we allow for parameters to be random we take two approaches: the one of Canova (1994) and the one of DeJong, Ingram and Whiteman (1996). In the first case empirical based distributions are constructed using existing estimates of these parameters or, when there are none, choosing a-priori an interval on the basis of theoretical considerations and imposing a uniform distribution on it. The distributions from which the parameters are drawn are displayed in the second column of Table 1.1. In the second case distributions for the parameters are assumed to be normal, with means equal to the basic calibrated parameters presented in column 1 while dispersions are a-priori chosen. The third column of Table 1.1 reports these distributions.

We generate samples of 95 observations to match the sample size of actual data. Because the initial conditions for the capital stock are set arbitrarily, the first 50 observations for each replication of the model are discarded. The number of replications used for each exercise is 500.

2.5.2 The Results

Table 2.2 summarizes the results obtained using four different evaluation approaches. Each row reports how the model fares in reproducing the spectral densities of saving and investment and the saving-investment coherence for US and Europe on average at business cycle frequencies (cycles of 3-8 years).

As a reference for comparison, the two first rows report the average spectral densities and coherences at business cycle frequencies for actual and simulated data when parameters are fixed (Kydland and Prescott approach). National saving is highly correlated with domestic investment but the average coherence at business cycle frequencies is higher for Europe than for the US. The variability of both US series is also higher and

US investment are almost two times more volatile than European ones. This pattern does not depend on the averaging procedure we choose; in fact, it is present at every frequency within the range we examine.

Given the symmetry of the model specification, the variability of simulated saving and investment is similar in both continental blocks, it is somewhat lower than the actual data for Europe, but definitively too low relative to the actual US series. Moreover, as in the actual european data, the variability is higher for national savings than for domestic investment. Consistent with Baxter and Crucini's claims, the model produces high national saving and investment correlations at business cycle frequencies. In fact, the model coherences for the US are higher than those found in the actual data.

The following rows of Table 2.2 check whether the above results persist when the performance of the model is evaluated using some of the procedures described in this chapter.

The first approach, which we use as a benchmark, is the one of Watson (1993). Given the spectral density matrix of the actual saving and investment for the two economic blocks, we calculate the spectral density matrix of the approximation error and compute the measure of fit (2.4) where Z includes frequencies corresponding to cycles of 3-8 years. Since in the model there are two technology disturbances, the spectral density matrix of simulated saving and investment for the two countries is singular and of rank equal to two. Therefore, to minimize the variance of the approximation error we consider two different identification schemes: in "identification 1" we jointly minimize the error term of the saving and investment of the first country (row 3 of Table 2.2) and in "identification 2" we jointly minimize the saving and investment errors of the second country (row 4 of Table 2.2). Note that to generate $R_j(\omega)$ we make two important assumptions: (i) that the spectral density matrix of the actual and simulated data can be measured without error and (ii) that the parameters of the model can be selected without error.

The results suggest that the fit of the model depends on the identification scheme used. On average, the size of the error at business cycle frequencies is between 2% and 5% of the actual spectral density of those variables whose variance is minimized

and between 20% and 30% of the actual spectral density of other variables, suggesting that "some" error should be added to the model to capture the features of the spectral density matrix of the data. Overall, we find small differences in the fit for the two continental blocks, and within continental blocks between the two variables of interest. Changes in the coherences across identifications are somewhat relevant and the model appears to fit coherences much better when we minimize the variance of US variables.

To show how the Monte Carlo techniques discussed in this paper can be used to evaluate the quality of the model's approximation to the data we compute three types of statistics. First, we report how many times on average, at business cycle frequencies, the diagonal elements of the spectral density matrix and the coherences of model generated data lie within a 95% confidence band for the corresponding statistics of actual data. That is, we report

$$T_1 = \int_{\omega_1}^{\omega_2} \int_{S_1(\omega)}^{S_2(\omega)} p_\omega(x) dx d\omega$$

where $S_1(\omega)$ and $S_2(\omega)$ are the lower and upper limits for the asymptotic 95% confidence band for the spectral density of actual data, ω_1 and ω_2 are the lower and upper limits for the range of business cycle frequencies and $p_\omega(x)$ is the empirical distribution of the simulated spectral density matrix for the four series at frequency ω.

If the spectral density matrix of the actual data is taken to be the object of interest to be replicated, T_1 reports the *confidence* of a test which assumes that the model is the correct DGP for the actual data. If we are not willing to assume that the model is the correct DGP for the actual data, these numbers judge the quality of the approximation by informally examining the magnitude of the probability coverings. No matter which interpretation we take, a number close to 95% would indicate a "good" model performance at a particular frequency band.

We compute 95% confidence bands for the actual data in two ways: using asymptotic distribution theory (as in the approach presented in the next chapter) and using a version of the parameteric bootstrap procedure of Diebold, Ohanian and Berkowitz (1995). In this latter case, we run a four variable VAR with 6 lags and a constant, construct replications for saving and investment for the two countries by bootstrapping the residuals of the VAR model, estimate the spectral density matrix of the data

for each replication and extract 95% confidence bands after ordering the replications, frequency by frequency.

Replications for the time series generated by the model are constructed using Monte Carlo techniques in three different ways. In the first case we simply randomize on the innovations of the exogenous technology process, keeping their distribution fixed (as in Gregory and Smith (1991)), and use the basic parameter setting displayed in the first column of Table 2.1. In the second and third cases parameters are random and drawn from the distributions listed in the second and third columns of Table 2.1. The results appear in rows 5 to 7 under the heading "Probability Covering". To economize on space and because simulated results are similar when the 95% confidence bands for actual data are computed asymptotically or by bootstrap, row 5 presents the percentage probability covering using an asymptotic 95% band when only the stochastic processes of the model are randomized, row 6 presents the probability covering using an asymptotic 95% band when we randomize on the exogenous stochastic processes of the model and parameters are drawn from empirically based distributions, and row 7 when parameters are drawn from normal prior distributions.

The results obtained with this testing approach highlight interesting features of simulated data. With fixed parameters, the average percentage of times the model spectra is inside the 95% band of the actual spectra is, in general, much smaller than 95%, its magnitude depends on the series and it is highest for European saving. When we randomize the parameters using DIW approach, results are more uniform across series and the probability covering is always of the order of 30% while when we randomize using empirical based distributions, the average percentage of times model's spectra are inside the 95% confidence band is somewhat lower. These results occur because with random parameters, simulated distributions are shifted and stretched: the model produces a wider range of variabilities than those possibly consistent with the data and this reduces the percentage of times simulated data are inside the asymptotic 95% band for each frequency. For coherences the results are very similar across the three rows: in this case, adding parameter variability does not change the outcomes. This is because parameter variability increases the volatility of saving and investment and their covariance by the same factor and this factor cancels out in the computation of

coherences. In general, we find that the model slightly "overfits" US coherences, i.e. on average too many simulations fall inside the asymptotic 95% band, while the opposite is true for European coherences. However, with empirical based priors, the coverage in both cases is close to 95%.

In sum, this evaluation procedure confirms that the model is better suited in matching coherences than volatilities at business cycle frequencies and that the covering properties of the model do not improve when we allow the parameters to be random.

To gain further evidence on the properties of the simulated distributions of the data, we next compute a second statistic: the percentile of the simulated distribution of the spectral density matrix of saving and investment for the two countries, where the value of the spectral density matrix of actual data (taken here to be estimated without an error) lies, on average, at business cycle frequencies. Implicitly, this p-value reports, on average over the selected frequency band, the proportion of replications for which the simulated data is less than the historical value. In other words, if $\tilde{S}_y(\omega)$ is the spectral density matrix of the actual data at frequency ω, we report

$$T_2 = \int_{\omega_1}^{\omega_2} \int_{-\infty}^{\tilde{S}_y(\omega)} p_\omega(x) dx d\omega$$

where all variables have been previously defined. Seen through these lenses the spectral density matrix of the actual data is treated as a "critical value" in examining the validity of the theoretical model. Values close to 0% (100%) indicate that the actual spectral density matrix is in the left (right) tail of the simulated distribution of the spectral density matrix of simulated data at that particular frequency band, in which case the model is poor in reproducing the statistics of interest. Values close to 50%, on the other hand, suggest that the actual spectral density matrix at those frequencies is close to the median of the distribution of the spectral density matrix of simulated data so the model is appropriate at those frequencies. Note also that values of the statistic in the range $[\alpha, 100 - \alpha]$, where α is a chosen confidence percentage, would indicate that the model is not significantly at odds with the data. We report the results of this exercise in rows 8 to 10 of Table 2.2 under the heading "Critical Value". Row 8 presents results when only the innovations of the technology disturbances are randomized, row

9 displays results when the parameters are drawn for normal priors and row 10 when parameters are drawn from an empirical based distribution.

As expected, the model with fixed parameters is unable to match the variabilities of the four series at business cycle frequencies. For all variables the statistics of actual data are in the right tail of the simulated distribution of the statistics at each frequency, i.e., a large proportions of simulations generate average values for the spectral density at business cycle frequencies which are lower than those found in the actual data. For European variables however, the picture is less dramatic. With parameter variability the picture changes. For all variables it is still true that actual variability exceeds the median of the simulated distribution on average at business cycle frequencies, but, at least with empirical priors, it is now within the interquartile range of the simulated distribution for three of the four variables. This is because parameter variability pushes the median of the simulated distribution close to the actual values. In essence, with parameter variability the model generates two features which improve its overall distributional fit: a wider range of variabilities at business cycle frequencies (with a somewhat larger percentage of more extreme values) and a less concentrated and less skewed distribution.

For coherences the results are somewhat different. With fixed parameters the model generates average coherences at business cycle frequencies which are much higher than in the data for the US but close to the median for Europe (actual values are in the 15th and 50th percentile). With random parameters (and empirical based priors), the situation improves for the US (actual coherence moves up to the 33rd percentile) but worsens for Europe. Once again, parameter variability enhances the range of possibilities of the model but it fails to tilt the distribution so as to more adequately reproduce the data.

Taken together, the results of these two exercises suggest that with fixed parameters the model generates a distribution for variability which is skewed to the left and only partially overlapping with a normal asymptotic range of variabilities for the data. For coherences the opposite occurs: the overlapping is high but also the skewness within the band is high. Parameter uncertainty, by tilting and stretching the shape of the simulated distribution, ameliorates the situation and in terms of the distributions of

certain statistics used, actual and simulated data are almost indistinguishable.

To complete the picture, we finally compute the distributional properties of the model approximation error by Monte Carlo methods, i.e. we compute the distribution of the error needed to match the spectral density matrix of the actual data given the model's simulated spectral density matrix. To compute the distributional properties of the log of the error, we draw, at each replication, parameters and innovations from the posterior distribution of the VAR representation of the actual data, construct time series of interest following the procedure of DeJong, Ingram and Whiteman (1996) and estimate the spectral density matrix of the four series. At each replication, we also draw parameters and innovations series from the distributions presented in Table 2.1, construct the spectral density matrix of simulated data and compute $S_u^i(\omega) = S_y^i(\omega) - S_x^i(\omega)$, i.e. the error in matching the spectral density matrix of the data at replication i. By drawing a large number of replications we can construct a nonparametric estimate of this distribution (using e.g. kernels) and compute moments and fractiles at each frequency. If the model is the correct DGP for the data, the distribution for this error would be degenerate at each frequency. Otherwise the features of this distribution (median value, skewness, kurtosis, etc.) may indicate what is missing from the model to capture the features of interest in the data. The last three rows in Table 2.2 present the median (across replications) of the average error across business cycle frequencies for the six statistics of interest under the heading "Error". Once again, we performed the calculations randomizing both on the stochastic processes of the model and the parameters of the model. Row 11 reports the results when parameters are fixed and rows 12 and 13 when the simulated time series incorporate uncertainty in both stochastic processes and parameters.

The results are quite similar in the three cases for the diagonal elements of the spectral density matrix. The model fails to generate enough variability at business cycle frequencies for US investments while for the other three variables the error is much smaller. The magnitude of the difference is, however, significant. For example for US savings and keeping parameters fixed, the error is about one-third of the actual variability at business cycle frequencies. The results for coherences depend on which of the two countries we consider. For US variables, the model generates systematically

higher coherences (negative spectral errors) while for Europe the opposite is true. Relatively speaking, these errors are of smaller magnitude than those obtained comparing spectra. Adding parameter variability as in DIW does not change the results too much. However, when parameters are drawn from empirical based priors, the model generates higher coherences in both cases.

2.5.3 What did we learn from the exercises?

Our exercise pointed out several important features of the model used by Baxter and Crucini (1993). As claimed by the authors, we find it generates high coherences between national saving and investment at business cycle frequencies which are of the same magnitude as the actual ones for European saving and investment. However, we also saw that the model tends to generate coherences which are uniformly higher than those observed in US data and this is true regardless of whether we used fixed or random parameters. In particular, we show that in only about 20% of the simulations is the simulated coherence smaller than the actual one and that there is tendency of the model to cluster saving and investment correlations in the vicinity of 1. Nevertheless, also in this case, the magnitude of the error is small. The model performance is worse when we try to account for the variability of saving and investment for the two continental blocks at business cycle frequencies. With fixed parameters, the simulated distribution at business cycle frequencies is skewed toward lower than actual values for all variables of interest and that the degree of overlap of simulated and actual distributions varies between 8 and 50%. Parameter variability helps but it does not represent a complete solution to the problem. This is clearly demonstrated by the size of the median value of the spectral error at business cycle frequencies which is sometimes larger than the error obtained with fixed parameters and always positive.

These results suggest that if one is interested in replicating the distributional properties of the statistics of the data (rather than their point estimates), it is necessary to respecify the model, at least for the US. What is primarily needed are two types of features. First, we need some real friction, maybe by adding a new sector (non-traded goods) which uses capital to produce goods; this modification is likely to reduce the median value of the distribution of correlation of saving and investment at business

cycle frequencies. Second, we need an additional propagation or variability enhancing device, maybe in the form of a lower adjustment cost of capital or higher elasticity of investment to technology innovations. For the US this can bring simulated variabilities at business cycle frequencies more in the range of the values we found in the data.

2.6 Conclusions

The task of this chapter was to illustrate how simulation techniques can be used to evaluate the quality of a model's approximation to the data, where the basic theoretical model design is one which fits into what we call a calibration exercise. In section 2.2 we first provide a definition of what calibration is and then describe in detail the steps needed to generate time series from the model and to select relevant statistics of actual and simulated data. In section 2.3 we overview four different formal evaluation approaches recently suggested in the literature, comparing and contrasting them on the basis of what type of variability they use to judge the closeness of the model's approximation to the data. In section 2.4 we describe how to undertake policy analysis with models which have been calibrated and evaluated along the lines discussed in the previous two sections. Section 2.5 presents a concrete example, borrowed from Baxter and Crucini (1993), where we design four different simulation-based statistics which allow us to shed some light on the quality of the model approximation to the data, in particular, whether the model is able to reproduce the main features of the spectral density matrix of saving and investment for the US and Europe at business cycle frequencies. We show that, consistent with Baxter and Crucini's claims, the model qualitatively produces a high coherence of saving and investment at business cycle frequencies in the two continental blocks but it also has the tendency to generate a highly skewed simulated distribution for the coherence of the two variables. We also show that the model is less successful in accounting for the volatility features of US and European saving and investment at business cycle frequencies and that taking into account parameter uncertainty helps in certain cases to bring the properties of simulated data closer to those of the actual data.

Overall, the example shows that simulation based evaluation techniques are very useful to judge the quality of the approximation of fully specified general equilibrium

models to the data and may uncover features of the model which are left hidden by more simple but more standard informal evaluation techniques.

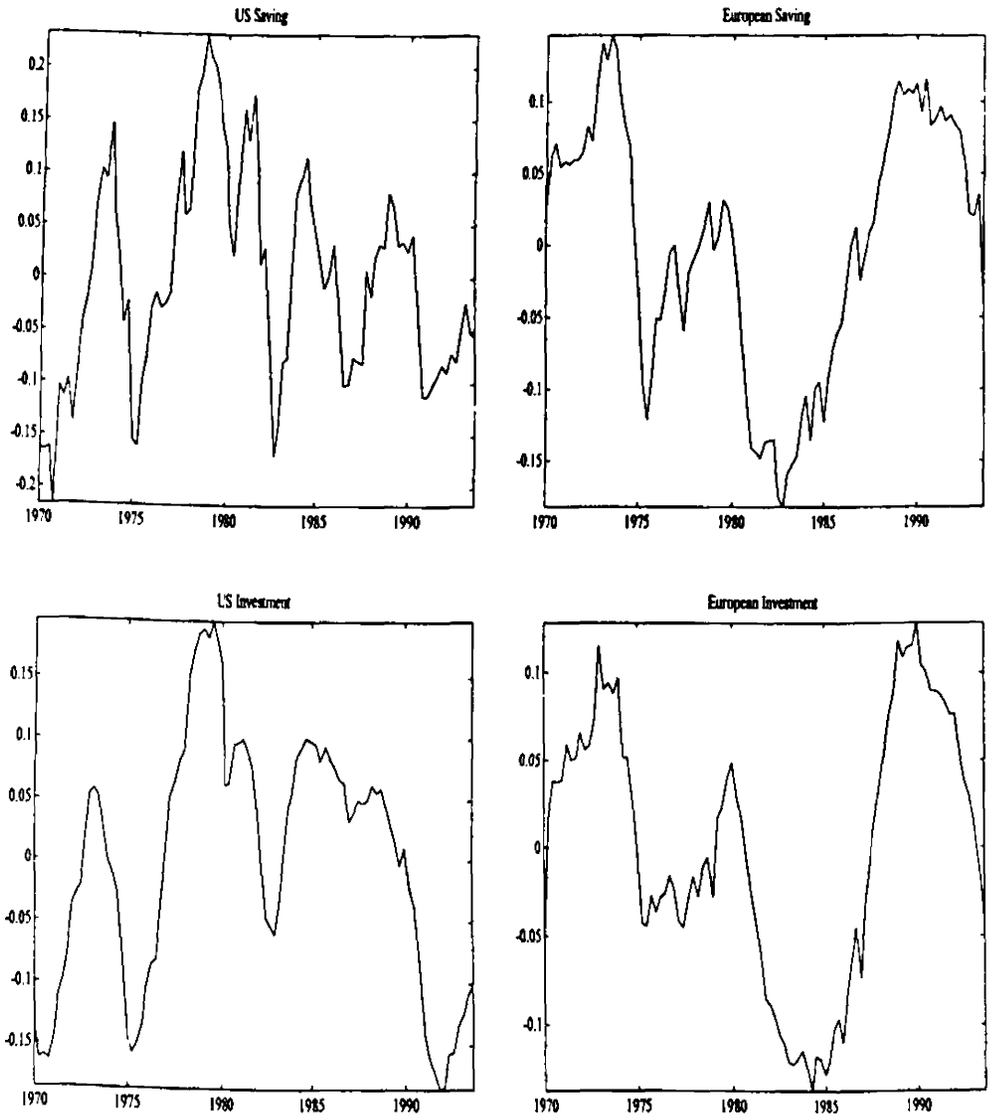

Figure 2.1: **US and European per capita saving and investment.** Linearly detrended logs of the series.

Table 2.1: **Parameter values used in the simulations**

Parameter	Basic	Empirical Density	Prior Normal
Steady State hours (H)	0.20	U [0.2, 0.35]	N (0.2, 0.02)
Discount Factor (β)	0.9875	Trunc.N [0.9855, 1.002]	N (0.9875, 0.01)
Risk Aversion (σ)	2.00	Trunc.$\chi^2(2)$[0, 10]	N (2, 1)
Share of Labor in Output (α)	0.58	U [0.50, 0.75]	N (0.58, 0.05)
Growth rate (θ_x)	1.004	N (1.004, 0.001)	1.004
Depreciation Rate of Capital (δ)	0.025	U [0.02, 0.03]	N (0.025, 0.01)
Persistence of Disturbances (ρ)	0.93	N (0.93, 0.02)	N (0.93, 0.025)
Lagged Spillover of Disturbances (ν)	0.05	N (0.05, 0.03)	N (0.05, 0.02)
Standard Deviation of Technology Innovations (σ_ϵ)	0.00852	Trunc.$\chi^2(1)$ [0, 0.0202]	N (0.00852, 0.004)
Contemporaneous Spillover (ψ)	0.40	N (0.35, 0.03)	N (0.4, 0.02)
Country Size (π)	0.50	U [0.10, 0.50]	0.5
Elasticity of marginal adjustment cost function ($\xi_{\phi'}$)	-0.075	-0.075	-0.075
Steady State Tobin's Q ($\frac{1}{\phi'}$)	1.0	1.0	1.0
Tax Rate (τ)	0.0	0.0	0.0

Notes: "Empirical density" refers to distributions for the parameters constructed using either existing estimates or a-priori intervals as in Canova (1994). "Prior Normal" refers to distributions for the parameters which are a-priori chosen as in DeJong, Ingram and Whiteman (1996). The range for the parameter is reported inside the brackets. The mean and the standard deviation for the distribution are reported inside the parentheses.

Table 2.2: **Fit of the model at Business Cycle frequencies**

	US Spectra		Europe Spectra		US Cohe	Europe Cohe
	S	I	S	I	S-I	S-I
Actual data	.0075	.0088	.0068	.0049	.85	.931
Simulated data	.0036	.0018	.0035	.0018	.94	.93
Watson approach						
Identification 1	.02	.05	.20	.23	.04	.13
Identification 2	.24	.21	.05	.04	.20	.15
Probability Covering						
Fixed parameters	46.46	8.63	55.71	43.57	98.99	92.91
Normal distribution	35.30	23.40	32.89	37.00	98.17	90.34
Empirical distribution	19.63	18.60	21.11	20.20	94.71	95.69
Critical Value						
Fixed parameters	90.80	99.89	82.16	93.91	15.60	49.04
Normal distribution	71.80	89.90	66.00	76.60	19.80	51.89
Empirical distributions	62.50	79.70	73.30	74.60	33.46	29.60
Error						
Fixed parameters	.0025	.0055	.0030	.0028	-.092	.004
Normal distribution	.0019	.0056	.0029	.0028	-.09	.008
Empirical distribution	.0013	.0058	.0042	.0035	-.061	-.029

Notes: Actual and simulated data are linearly detrended and logged, in real per capita terms. Simulations are undertaken using 500 draws. "Watson approach" reports the average statistic (2.4) at business cycle frequencies, "Probability covering" reports the average percentage covering at business cycle frequencies of the theoretical 95% range, "Critical value" the percentile where the actual data lies on average at business cycle frequencies, and "Error" the median error across simulations on average at business cycle frequencies. S refers to saving and I to investment.

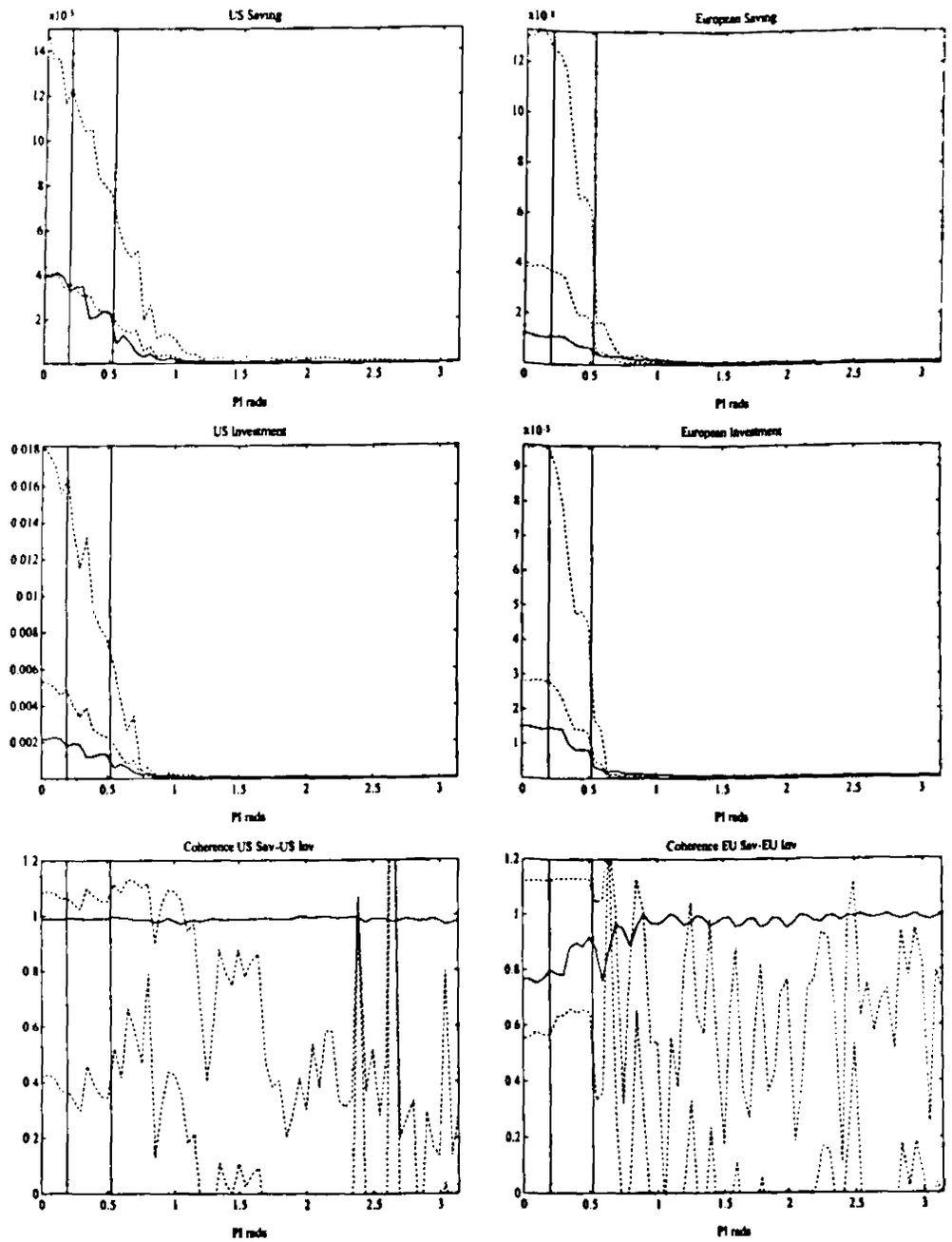

Figure 2.2: **Spectra and coherences of US and European per capita saving and investment** (linearly detrended logs of the series). 95% asymptotic confidence interval of estimated spectral densities and coherences for actual data displayed in dashed lines, spectra and coherences of simulated data for one draw in solid lines. Vertical lines indicate the frequencies associated to cycles of 8 and 3 years.

Chapter 3

A New Methodology for Assessing the Fit of Multivariate Dynamic Models

3.1 Introduction

The increasing complexity of the issues economists want to address has induced the wide use of multivariate dynamic models. However, their own complexity often implies the inability to obtain analytic solutions to these models and hence simulation techniques are used to approximate the equilibrium solution; i.e., simulate solution paths for the endogenous variables in terms of the exogenous variables and the parameters (see Marcet (1994) for a review of the application of simulation methods to economics). The empirical analysis of such models has to deal with the obvious misspecification not only when selecting the specific functionals linking the endogenous to the exogenous variables of the model, and when parameterizing the model's structure and the distribution of the exogenous processes, but also when finding an approximate solution. Precisely because of the various possible misspecifications, it is an important issue to assess adequately the ability of a simulated model to reproduce certain aspects observed in actual data (i.e. assess the fit of the model) as well as to compare the performance of alternative models.

The many possible sources of misspecification have led to the widespread adoption

of the informal calibration technique[1] pioneered by Kydland and Prescott (1982). See Chapter 1 for a detailed explanation of the calibration approach. Typically, a model is specified given a concrete question a researcher wants to study. The model is then solved (usually through an approximation method), the parameters are given fixed values and exogenous processes fixed distributions and data series for the variables of interest are generated by simulation of the model. The assessment of the performance of a model is typically reduced to a relatively subjective comparison of two reduced sets of summary statistics obtained from the simulated and the actual data. The model economy is considered a "good" approximation of the actual world if it can broadly reproduce the observed features of the series that it purports to model. The adequacy of a particular parameterization is typically checked through sensitivity analysis, which essentially consists of computing and comparing the same statistics for different parameterizations. Comparison of competing models very seldom takes place and when it does it is typically reduced to a similarly informal subjective comparison of selected statistics. These procedures, based on comparing the similarities between simulated and actual data are essentially ad-hoc, lack statistical foundations and ignore information that could be used for model evaluation purposes.

Recent research in applied macroeconomics and time series econometrics has suggested alternatives to such informal approach to assess the fit of a model. In Chapter 1 we roughly classify those alternatives into four categories (see also Kim and Pagan (1994) for a review of recent methods for evaluating calibrated models). This paper contributes to this literature by proposing an alternative measure of fit to evaluate a multivariate dynamic model and to compare the performance of alternative model specifications. The evaluation methodology proposed here addresses two important issues highlighted in the literature in a unifying fashion.

Firstly, we explicitly acknowledge that the solution paths generated by the model for the variables of interest are only approximations to the true model solution. Some simulation techniques approximate model solutions with an arbitrary degree of accu-

[1] Economic models tend to be small and strongly theory based, and hence, as Pagan (1994) stresses, unlikely to obey the 'axiom of correct specification'. That is the idea underlying all calibration exercises, as made specific by Kydland and Prescott (1991): "..., no attempt is made (in the Calibration Approach) to determine the true model. All model economies are abstractions and are by definition false".

racy but are so demanding in terms of either complexity or computer time that the most common position is to use faster but less accurate approximations (see Marcet (1994)) [2]. Hence, it may not be so reasonable to assume that the approximation is the original model as is usual. Watson (1993) also recognises that there is an approximation error but, contrary to his approach, we take it into account when deriving a formal test of the distance between the model and the observed data.

Secondly, as in the last approach, our tests take into account both the sampling variability of actual data and the uncertainty in the simulated series. While not excluding the possibility of stochastic parameters in the model, the uncertainty we consider in the model derives from the fact that there exists an approximation error. As in Diebold, Ohanian and Berkowitz (1995), the measure of distance and tests presented in this paper evaluate how well the model matches the spectral density matrix of the actual data. But hey assume that model statistics can be estimated without error simply by simulating very long time series from the model and hence use only the sampling variability of actual data statistics (evaluated with bootstrap algorithms) to propose goodness-of-fit criteria and to derive associated optimal parameter estimators. Although they deal with the issue of parameter uncertainty, this approach does not take into account the possible misspecification induced when approximating the solution of the model. Instead, we compare actual to simulated data by treating them as samples from their unknown DGP and hence both spectral density matrices are estimated with error. The required asymptotic theory is developed to test the hypothesis that the multivariate spectral density matrix of the model and the actual data (or of two models) are alike (either equal or differing to an arbitrary prespecified limit). DeJong, Ingram and Whiteman (1996) and Canova and De Nicoló (1995) suggest also measures of fit which are symmetric in the statistical treatment of model and actual data. The main differences between the moethodology proposed in this Chapter and that of DeJong, Ingram and Whiteman (1996) and of Canova and De Nicoló (1995) are, first, that we estimate both sets of statistics in a classical way instead of using Bayesian methods

[2]Den Haan and Marcet (1994) propose a simple test for the accuracy of the numerical approximation to the solution of a model. Such a test is certainly helpful in selecting more accurate approximations for a given model, but as long as the approximation error exists it may affect the properties of the model and how well it reproduces the observed properties of the actual data.

and, second, that model and actual data (or another model) statistics are compared using asymptotic tests.

Section 3.2 proposes a formal *measure of fit* to evaluate models against actual data using multivariate frequency domain techniques. An asymptotic test is derived for the hypothesis that the distance between the spectral density matrices of simulated model series and actual data is zero or less than an arbitrary prespecified bound. It is especially suitable for assessing the performance of models that focus on the dynamic behavior of a set of key variables at a certain frequency range, such as business cycle models. In a similar fashion, section 3.3 derives a formal test for the equivalence of competing models, possibly misspecified, or of alternative parameterizations of a same model (i.e. to perform global sensitivity analysis). It can be seen as a *comparison test* between different model specifications. The test is able to address the complicated and interesting issue of comparing misspecified models, testing whether they are similar to each other while being different from the actual DGP. Section 3.4 examines the finite sample properties of both tests via Monte Carlo experiments. The sensitivity of the tests proposed in this paper to the sample size and to the parameter structure is also studied. Section 3.5 applies the *fit* test to alternative versions of an International Real Business Cycle based on Backus, Kehoe and Kydland (1993). We want to evaluate the effect of final goods trade, common shocks and spillovers across national disturbances on the macroeconomic interdependencies between countries. Assessment of the fit of these models focusses on how well they reproduce the bivariate spectral density matrix of the US and European GDPs at business cycle frequencies using the *fit* test. The spectral density matrices implied by each alternative model are compared using the *comparison* test proposed in Section 3.3. Section 3.6 summarizes and concludes.

3.2 A measure of distance between simulated and actual data

We are interested in comparing the dynamic properties of a multivariate vector of observed economic data with those generated from a simulated multivariate dynamic model.

Let y_t be the N×1 vector of actual data series and x_t be the N×1 vector of data simulated from the model, where t = 1,...,T. We view the artificial series x_t in a similar way as the observed data y_t, as samples of an unknown DGP. The reason being that x_t is not typically generated through an analytical solution of the model but approximating that solution with some simulation algorithm. The DGP of the artificial data is unknown but is known to be sufficiently close to the model (see Marcet (1994)), so that the artificial series obtained can be used to evaluate the performance of the theoretical model as long as we take the possible approximation error into account.

In order to take into account all of the interactions between actual and artificial data, we use the joint spectral density matrix for the 2N×1 vector $z_t = [\, y_t' \;\; x_t' \,]'$. For each frequency $\omega \in [-\pi, \pi]$, let $f(\omega;\gamma) = \{f_{ij}(\omega;\gamma)\}$ denote the 2N×2N theoretical spectral density matrix of vector z_t in which the model series x_t are obtained using the parameter vector γ. The ij-th element represents the corresponding crosspectrum between a pair of variables, z_{it} and z_{jt}, i,j = 1,...,2N. $f(\omega;\gamma)$ is arranged as follows

$$f(\omega;\gamma) = \begin{pmatrix} f^{\mathbf{y}}(\omega) & f^{\mathbf{yx}}(\omega;\gamma) \\ f^{\mathbf{xy}}(\omega;\gamma) & f^{\mathbf{x}}(\omega;\gamma) \end{pmatrix}$$

where $f^{\mathbf{y}}(\omega)$ ($f^{\mathbf{x}}(\omega;\gamma)$) corresponds to the crosspectra between the actual (artificial) series and the other two submatrices correspond to crosspectra between pairs of actual and model series. Let $\hat{f}(\omega;\gamma) = \{\hat{f}_{ij}(\omega;\gamma)\}$ denote the estimated spectral density matrix, defined as

$$\hat{f}(\omega;\gamma) = \frac{1}{2\pi} \sum_{\tau=-T+1}^{T-1} k_M(\tau)\hat{\Gamma}(\tau;\gamma)e^{-i\omega\tau}$$

where $\hat{\Gamma}(\tau;\gamma)$ is the variance-covariance matrix estimate of vector z_t for lag τ, $k_M(\tau)$ is the lag window function and M is the lag/spectral window parameter. Under general conditions of stationarity of z_t, and standard assumptions on $k_M(\tau)$ and M (see Appendix 1 for the assumptions and the properties of spectral estimators), $\hat{f}(\omega;\gamma)$ is a consistent and asymptotically unbiased estimator of $f(\omega;\gamma)$ with the following asymptotic distribution

$$\sqrt{\frac{\nu}{2}}\mathrm{vec}\hat{f}(\omega;\gamma) \;\dot{\sim}\; \mathrm{CN}_{N^2}\left(\sqrt{\frac{\nu}{2}}\mathrm{vec}f(\omega;\gamma), \overline{f(\omega;\gamma)} \otimes f(\omega;\gamma)\right) \quad for\ \omega \neq 0, \pm\pi \qquad (3.1)$$

where $\dot{\sim}\ \mathrm{CN}_{N^2}$ indicates an asymptotic multivariate complex Normal distribution of dimension N^2, $\otimes$ denotes the kronecker product and ν is a constant called "equivalent

degrees of freedom" of the spectral estimator and is defined as $\nu = \frac{2T}{M \int_{\tau=-\infty}^{+\infty} k_M^2(\tau)d\tau}$ (the value of ν for each lag window function is tabulated, see Priestley (1981)).

We define, for each frequency ω, the *theoretical distance* between the model and the observed data by:

$$D(\omega;\gamma) = \mathrm{S}\ \mathrm{vec}f(\omega;\gamma) = \mathrm{vec}f^{\mathrm{y}}(\omega) - \mathrm{vec}f^{\mathrm{x}}(\omega;\gamma) \tag{3.2}$$

where vec(.) is the column vectorization operator, and S is a $N^2 \times (2N)^2$ selection matrix that transforms $\mathrm{vec}f(\omega;\gamma)$ into the difference between the elements of its submatrices f^{y} and f^{x}. Note that $D(\omega;\gamma)$ results in a $N^2 \times 1$ vector.

As an illustration of our measure of distance $D(\omega;\gamma)$, let y_t follow a bi-variate zero-mean VAR(1) process, $y_t = \Phi^{\mathrm{y}} y_{t-1} + \epsilon_t$, t = 1,...,T, where Φ^{y} is a 2×2 parameter matrix whose eigenvalues lie all inside the unit circle, and ϵ_t is a bi-variate white noise (WN) vector, with $E[\epsilon_t\, \epsilon_t'] = \Omega_\epsilon$. Then, we can express y_t in its MA(∞) form

$$y_t = (I_2 - \Phi^{\mathrm{y}} L)^{-1}\epsilon_t = \sum_{\tau=0}^{\infty}(\Phi^{\mathrm{y}})^{\tau}\epsilon_{t-\tau} = \sum_{\tau=0}^{\infty}(\Phi^{\mathrm{y}} L)^{\tau}\epsilon_t = \sum_{\tau=0}^{\infty} A_\tau^{\mathrm{y}} L^{\tau}\epsilon_t = A^{\mathrm{y}}(L)\epsilon_t$$

and obtain the theoretical spectrum:

$$f^{\mathrm{y}}(\omega) = \frac{1}{2\pi}(I_2 - \Phi^{\mathrm{y}} e^{-i\omega})^{-1}\Omega_\epsilon(I_2 - (\Phi^{\mathrm{y}})' e^{i\omega})^{-1} = A^{\mathrm{y}}(e^{-i\omega}) f^{\epsilon}(\omega) A^{\mathrm{y}}(e^{i\omega})'$$

The simulated series generated from the model x_t being also covariance-stationary, they can be expressed in their MA form, $x_t = A^{\mathrm{x}}(L)\, u_t$, where u_t is a bi-variate WN process as ϵ_t. Hence,

$$f^{\mathrm{x}}(\omega;\gamma) = A^{\mathrm{x}}(e^{-i\omega}) f^{\mathrm{u}}(\omega;\gamma) A^{\mathrm{x}}(e^{i\omega})'$$

A plot of the row i, column j element of A_τ as a function of the lag τ is called the *impulse response function.* Hence, the measure of distance between model and actual data series $D(\omega;\gamma)$ we have defined as the difference between their spectral density matrices, can also be thought of as a measure of distance between the theoretical *impulse responses* of the model and the actual data,

$$D(\omega;\gamma) = \mathrm{vec}f^{\mathrm{y}}(\omega) - \mathrm{vec}f^{\mathrm{x}}(\omega;\gamma) = \mathrm{vec}[A^{\mathrm{y}}(e^{-i\omega}) f^{\epsilon}(\omega) A^{\mathrm{y}}(e^{i\omega})' - \tag{3.3}$$
$$A^{\mathrm{x}}(e^{-i\omega}) f^{\mathrm{u}}(\omega;\gamma) A^{\mathrm{x}}(e^{i\omega})'] = \frac{1}{2\pi}\mathrm{vec}[A^{\mathrm{y}}(e^{-i\omega})\Omega_\epsilon A^{\mathrm{y}}(e^{i\omega})' - A^{\mathrm{x}}(e^{-i\omega})\Omega_u A^{\mathrm{x}}(e^{i\omega})']$$

Since ϵ_t and u_t are WN processes, $f^{\epsilon}(\omega)$ and $f^{u}(\omega;\gamma)$ are flat, and equal if $\Omega_\epsilon = \Omega_u$. In that case, $D(\omega;\gamma)$ measures the distance frequency by frequency between the "squared" theoretical impulse responses of the actual data and those of the model. If $\Omega_\epsilon \neq \Omega_u$, the distance between the two spectral density matrices takes into account the different covariance structure of both actual and simulated data innovations sets i.e. ϵ_t and u_t.

Now we define the ***estimated distance*** between model and data by:

$$\hat{D}(\omega;\gamma) = S\text{vec}\hat{f}(\omega;\gamma) = \text{vec}\hat{f}^{y}(\omega) - \text{vec}\hat{f}^{x}(\omega;\gamma) \tag{3.4}$$

The asymptotic distribution and properties of $\hat{D}(\omega;\gamma)$ are derived from those of $\hat{f}(\omega;\gamma)$ (see Appendix 1):

(a) asymptotic complex Normal distribution

$$\sqrt{\frac{\nu}{2}}\left(\hat{D}(\omega;\gamma) - D(\omega;\gamma)\right) = \sqrt{\frac{\nu}{2}}S\text{vec}\left(\hat{f}(\omega;\gamma) - f(\omega;\gamma)\right) \dot{\sim} \text{CN}_{N^2}, \quad \omega \neq 0, \pm\pi \tag{3.5}$$

(b) asymptotic unbiasedness

$$\lim_{T\to\infty} E[\hat{D}(\omega;\gamma)] = \lim_{T\to\infty} E[S\text{vec}\hat{f}(\omega;\gamma)] = S\text{vec}f(\omega;\gamma) = D(\omega;\gamma) \tag{3.6}$$

(c)-(d) asymptotic variance-covariance structure

$$\Sigma_D(\omega;\gamma) = \lim_{T\to\infty} \text{var}\left[\sqrt{\frac{\nu}{2}}\hat{D}(\omega;\gamma)\right] = \lim_{T\to\infty} var[S\sqrt{\frac{\nu}{2}}\text{vec}\hat{f}(\omega;\gamma)] = \\ = S\overline{f(\omega;\gamma)} \otimes f(\omega;\gamma)S', \quad \omega \neq 0, \pm\pi \tag{3.7}$$

Therefore, for $\omega \neq 0, \pm\pi$,

$$\sqrt{\frac{\nu}{2}}\hat{D}(\omega;\gamma) \dot{\sim} \text{CN}_{N^2}\left(\sqrt{\frac{\nu}{2}}D(\omega;\gamma),\ \Sigma_D(\omega;\gamma)\right) \tag{3.8}$$

Recall that we are interested in evaluating the performance of multivariate dynamic models that are, in general, solved by approximation through simulation techniques. It means that the model yields a multivariate vector series of same dimension as the actual data for each time it is simulated, and hence a $\hat{f}_h(\omega;\gamma)$ is estimated keeping y_t fixed and using x_{ht} at each replication, for h = 1,...,H. In practice, what we are interested in obtaining is the average across the H replications of the estimated distance $\hat{D}(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H}\hat{D}_h(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H} S\ \text{vec}\hat{f}_h(\omega;\gamma)$. Given that x_{ht} are iid, that

average is the sample mean of iid random variables, where the sample size is H and the iid r.v. are $\text{vec}\hat{f}_h(\omega;\gamma)$ premultiplied by S, for h = 1,...,H. Hence, the sample mean across replications has same distribution and theoretical mean as each of its elements, $\text{vec}\hat{f}_h(\omega;\gamma)$, and a variance which is $\frac{1}{H}\text{var}(\text{vec}\hat{f}_h(\omega;\gamma))$. Hence, instead of (3.8), for H finite, the asymptotic distribution of $\hat{D}(\omega;\gamma)$ is:

$$\sqrt{H}\sqrt{\frac{\nu}{2}}\hat{D}(\omega;\gamma) \stackrel{\sim}{\sim} \text{CN}_{N^2}\Big(\sqrt{H}\sqrt{\frac{\nu}{2}}D(\omega;\gamma),\ S\overline{f(\omega;\gamma)}\otimes f(\omega;\gamma)S'\Big), \qquad (3.9)$$

In what follows, for simplicity, we will consider only the case of H=1. All results can be easily generalised to a generic H. However, it can be the case that for very short sample sizes where the convergence of $\hat{D}(\omega;\gamma)$ to its limiting distribution is not ensured and may be closer to a χ^2 distribution (see property (a) in Appendix 1), the actual number of $\hat{D}_h(\omega;\gamma)$ elements it aggegates (H) may matter, probably increasing the degrees of freedom of the χ^2 distribution. Obviously, this effect will only be noticeable for large H[3].

Next we present a test as to whether the elements of $D(\omega;\gamma)$ are significantly different from zero for a particular single frequency, without having to deal with complex distributions, and extend it to the case for when we are interested in the significance of the distance of a model from the observed data over a particular frequency range.

3.2.1 Assessing the fit of a model

Testing for the significance of the distance of a model from the actual data at a given frequency ω means that we are testing the following null hypothesis:

$$H_0 : \Lambda D(\omega;\gamma) = 0$$

where Λ is an $N^2\times N^2$ diagonal selection matrix. Such a matrix may have unequal diagonal elements so as to introduce weights in the measure of distance $D(\omega;\gamma)$ if we care about some relationship more than others. The ((i-1)×N+j)-th element of its diagonal represents the weight given to the relationship between the i-th and j-th elements in y_t and x_t, with i, j = 1,..., N.

[3]In fact, in Chapter 3 we find such effects when assessing versions of the Real Business Cycle model of King, Plosser and Rebelo (1988) using H=1000 with sample size 127.

To test H_0, we then construct the following test statistic

$$fit(\omega;\gamma) = \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right)}'\left(\Lambda\Sigma_D(\omega;\gamma)\Lambda'\right)^{-1}\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right) \tag{3.10}$$

which is a real number because we are multiplying element by element the standarized estimated distance vector of interest $(\Lambda\Sigma_D(\omega;\gamma)\Lambda')^{-\frac{1}{2}}\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)$ by its conjugate. Given that $\hat{f}(\omega;\gamma)$ is a consistent estimator of $f(\omega;\gamma)$, $\Sigma_D(\omega;\gamma)$ can be replaced by $\hat{\Sigma}_D(\omega;\gamma) = S\overline{\hat{f}(\omega;\gamma)} \otimes \hat{f}(\omega;\gamma)S'$. Thus, the test statistic $fit(\omega;\gamma)$ becomes

$$fit(\omega;\gamma) = \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right)}'\left(\Lambda\hat{\Sigma}_D(\omega;\gamma)\Lambda'\right)^{-1}\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)$$

Under H_0, (3.8) becomes $\sqrt{\frac{\nu}{2}}\hat{D}(\omega;\gamma) \dot{\sim} \mathrm{CN}_{N^2}\left(0, \Sigma_D(\omega;\gamma)\right)$ and, hence, the asymptotic distribution of the test statistic is

$$fit(\omega;\gamma) \;\dot{\sim}\; \chi^2_{(N^2-Q)} \tag{3.11}$$

where Q is the number of zero elements in the diagonal of Λ.

H_0 will be rejected and the distance between the model and the actual data found significantly different from zero if $fit(\omega;\gamma)$ is greater than the critical value of a $\chi^2_{(N^2-Q)}$, for a selected significance level α.

When we suspect a model to be false, we may be more interested in testing whether its distance to the actual data is smaller than an arbitrary constant C rather than testing the above null of zero distance. Then, the relevant null hypothesis is

$$H_0 : \Lambda D(\omega;\gamma) \;\leq\; C, \quad \forall\omega \in [\omega_1, \omega_2]$$

The only difference with respect to the test described above is that, under the null, the test statistic $fit(\omega;\gamma)$ has a non-central asymptotic $\chi^2_{(N^2-Q,\delta)}$ distribution, where δ is a non-centrality parameter of value $\delta = \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda D(\omega;\gamma)\right)}'\left(\Lambda\Sigma_D(\omega;\gamma)\Lambda'\right)^{-1}\left(\sqrt{\frac{\nu}{2}}\Lambda D(\omega;\gamma)\right)$.

We may want to test the significance of the distance of the model to the actual data for a given set of L frequencies $[\omega_1, \omega_2]$, where $\omega_1, \omega_2 \neq 0, \pm\pi$, e.g. the frequencies associated with the business cycle, which is typically associated to those cycles whose periods lie within 2 and 8 years. Then, the null hypothesis is

$$H_0 : \Lambda D(\omega;\gamma) = 0, \quad \forall\omega \in [\omega_1, \omega_2]$$

Under H_0, the test statistic $fit(\omega;\gamma)$ becomes

$$\text{fit}([\omega_1,\omega_2];\gamma) = \sum_{\omega=\omega_1}^{\omega_2} \left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right)' \left(\Lambda\hat{\Sigma}_D(\omega;\gamma)\Lambda'\right)^{-1} \sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma) \sim \chi^2_{L(N^2-Q)} \quad (3.12)$$

since it is the sum of L independent $\chi^2_{(N^2-Q)}$ variates.

3.3 Comparing alternative models

An important characteristic of a model is its performance relative to other models in capturing a particular aspect of reality. In this section we develop a formal test for assessing whether *two different model specifications* display similar dynamic properties for a selected group of variables. If this is the case, they can be considered equally successful in capturing the dynamic properties observed in the actual data. How far or close to the actual data each of the models is can be assessed using the *fit* test presented in the previous section.

Let $x_t^i(\gamma_i)$ denote the N×1 vector obtained by simulating model m_i with the particular set of parameters γ_i. We are interested in taking into account all the interactions between the two alternative model specifications $(m_1;\gamma_1)$ and $(m_2;\gamma_2)$ we want to compare. Therefore, the relevant z_t vector whose spectral density matrix we want to estimate is the 2N×1 vector $z_t = [\, x_t^1(\gamma_1)' \;\; x_t^2(\gamma_2)']'$. Both $f(\omega;\gamma_1,\gamma_2)$ and $\hat{f}(\omega;\gamma_1,\gamma_2)$ are 2N×2N matrices.

Alternatively, we could also compare the relative success of the same model m_1 under two alternative parameterizations γ_1 and γ_2. Then, the relevant z_t vector would be $z_t = [\, x_t^1(\gamma_1)' \;\; x_t^1(\gamma_2)']'$. This can be regarded as a formalization of the sensitivity analysis the researcher may want to undertake over certain elements of the parameter vector γ. A more global sensitivity analysis could be performed with a modified version of the *fit* test, in which y_t is compared to an average of x_t^i over γ. If one considers all possible realistic values of the parameter space we would be in fact introducing a way to deal with parameter uncertainty, along the same lines as Canova (1994)-(1995), DeJong, Ingram and Whiteman (1996) and Canova and De Nicoló (1995).

In a similar way as in Section 3.2, we define for each frequency ω the *theoretical*

distance between two alternative model specifications (m_1, γ_1) and (m_2, γ_2) by:

$$D(\omega;\gamma_1,\gamma_2) = Svecf(\omega;\gamma_1,\gamma_2) = vecf^2(\omega;\gamma_2) - vecf^1(\omega;\gamma_1) \qquad (3.13)$$

where, as in section 3.2, S is a $N^2 \times (2N)^2$ selection matrix and $D(\omega;\gamma_1,\gamma_2)$ results in a $N^2 \times 1$ vector. Note that we compare the dynamic properties of two models independently of how close each of them is to the actual data. Had we included the actual data vector y_t, we would have defined $D(\omega;\gamma_1,\gamma_2)$ as the difference between the fit of each model, i.e. $D(\omega;\gamma_1,\gamma_2)$ would have been defined as $(vecf^2(\omega;\gamma_2) - vecf^y(\omega)) - (vecf^1(\omega;\gamma_2) - vecf^y(\omega))$, which is equal to (3.13).

Our definition of the distance between two alternative models allows us to test the null that two models have the same spectral density matrices (or submatrices if we are interested in the dynamic properties of only a subset of the variables included in the models) both in the case in which one of the models is the actual DGP and when neither of them is, i.e. when the comparison is made between misspecified models.

The *estimated distance* between the two alternative model specifications is defined as follows:

$$\hat{D}(\omega;\gamma_1,\gamma_2) = S\hat{vecf}(\omega;\gamma_1,\gamma_2) = \hat{vecf}^2(\omega;\gamma_2) - \hat{vecf}^1(\omega;\gamma_1) \qquad (3.14)$$

and has similar asymptotic properties to $\hat{D}(\omega;\gamma)$. For $\omega \neq 0, \pm\pi$,

$$\sqrt{\frac{\nu}{2}}\hat{D}(\omega;\gamma_1,\gamma_2) \dot{\sim} \mathrm{CN}_{N^2}\Big(\sqrt{\frac{\nu}{2}}D(\omega;\gamma_1,\gamma_2),\ \Sigma_D(\omega;\gamma_1,\gamma_2) = S\overline{f(\omega;\gamma_1,\gamma_2)}\otimes f(\omega;\gamma_1,\gamma_2)S'\Big) \qquad (3.15)$$

Note that we are assuming that both model specifications have been simulated the same number of times, H: As before, for values of H$\neq$1, $\hat{D}(\omega;\gamma_1,\gamma_2)$ is replaced by $\sqrt{H}\hat{D}(\omega;\gamma_1,\gamma_2)$.

Testing for the equal performance of two model specifications with respect to the dynamic properties of a selected set of series requires testing whether the null hypothesis

$$H_0 : \Lambda D(\omega;\gamma_1,\gamma_2) = 0$$

at frequency ω, is accepted; again, Λ is a selection matrix.

In a similar fashion as the $fit(\omega;\gamma)$ test, we construct the following test statistic:

$$\mathrm{comp}(\omega;\gamma_1,\gamma_2) = \overline{\Big(\sqrt{\frac{\nu}{2}}\Lambda\hat{\mathrm{D}}(\omega;\gamma_1,\gamma_2)\Big)}'\Big(\Lambda\hat{\Sigma}_{\mathrm{D}}(\omega;\gamma_1,\gamma_2)\Lambda'\Big)^{-1}\sqrt{\frac{\nu}{2}}\Lambda\hat{\mathrm{D}}(\omega;\gamma_1,\gamma_2) \qquad (3.16)$$

Under the null,

$$comp(\omega;\gamma_1,\gamma_2) \dot{\sim} \chi^2_{(N^2-Q)} \tag{3.17}$$

H_0 will be rejected and the relative distance between the model specifications is significantly different from zero if $comp(\omega;\gamma_1,\gamma_2)$ is greater than the critical value of a $\chi^2_{(N^2-Q)}$, for a given significance level α, and where Q is the number of zero elements in the diagonal of Λ.

Here, too, we may want to test the significance of the distance between two alternative model specifications for a given set of L frequencies, $[\omega_1,\omega_2]$, where $\omega_1,\omega_2 \neq 0,\pm\pi$. Then the null hypothesis is

$$H_0 : \Lambda D(\omega;\gamma_1,\gamma_2) = 0, \quad \forall\omega \in [\omega_1,\omega_2].$$

To test such H_0, we use an aggregated version of $comp(\omega;\gamma)$

$$\text{comp}([\omega_1,\omega_2];\gamma_1,\gamma_2) = \sum_{\omega=\omega_1}^{\omega_2} \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma_1,\gamma_2)\right)}' \left(\Lambda\hat{\Sigma}_D(\omega;\gamma_1,\gamma_2)\Lambda'\right)^{-1} \sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma_1,\gamma_2) \tag{3.18}$$

which has a $\chi^2_{L(N^2-Q)}$ asymptotic distribution under H_0.

3.4 Performance of the tests: Monte Carlo evidence

In this section we present some Monte Carlo evidence to examine the finite sample properties of the two proposed tests, $fit([\omega_1,\omega_2];\gamma)$, and $comp([\omega_1,\omega_2];\gamma_1,\gamma_2)$ for the case of a bivariate model (N=2). Experiments have been conducted for the following sample sizes[4] : T = 100, 200 and 500. In all cases, we evaluate the performance of the bivariate model at business cycle frequencies and we define $[\omega_1,\omega_2]$ as the set of frequencies associated with cycles 8 to 2 years long, and all variables are given equal weight i.e. $\Lambda = I_{N^2} = I_4$.

The y_t series have been generated from a bivariate VAR(1) of the form:

$$y_t = \Phi^y y_{t-1} + \epsilon_t, \quad t = 1,...,T$$

[4]We have also experimented with T=1000 obtaining results very similar to those with T=500, therefore we do not report them. There are slight differences but all in the same lines as the changes observed when going from T=200 to T=500.

where Φ^y is a 2x2 parameter matrix. In particular, $\Phi^y = \begin{pmatrix} 0.7 & 0.1 \\ 0.2 & 0.6 \end{pmatrix}$. These concrete values generate a bi-variate covariance stationary process in which the two series are correlated to each other and both show the "typical spectral shape" Granger (1964) attributes to macroeconomic data series, i.e. most of the variability concentrated in the lower frequencies.

In order to see the sensitivity of the tests to the particular dynamic structure of the model, the x_t^i series have been simulated from three alternative models, i = 1, 2 and 3, of the form:

$$x_t^i = \Phi^i x_{t-1}^i + u_t^i, \quad t = 1, ..., T$$

where Φ^i is a 2×2 parameter matrix and the residuals vector, u_t^i, is a bivariate white noise (WN) process. The three models simulated to obtain the x_t^i vector series differ on their parameter structure.

- Model 1: VAR(1) with $\Phi^1 = \Phi^y$.
- Model 2: No spillovers: VAR(1) with $\Phi^2 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.6 \end{pmatrix}$.
- Model 3: No dependence structure: Bivariate WN process, $x_t^3 = u_t^3$. Hence, $\Phi^3 = 0_2$.

Error vectors, both u_t^i and ϵ_t, have been generated with the Matlab 4.2 random generator from a standard bivariate Normal distribution[5]. Since $\{u_t^i\}_{t=1}^T$ and $\{\epsilon_t\}_{t=1}^T$ have the same spectral density matrices, $\{x_t^i\}_{t=1}^T$ will have equal spectral density matrix to $\{y_t\}_{t=1}^T$ as long as the parameters of their respective DGPs are the same. 100 extra observations were generated for each error $\{\epsilon_t\}_{t=1}^T$ and residual $\{u_t\}_{t=1}^T$ vector sequences in order to avoid initial condition problems.

Before proceeding, we have to choose both the functional form of the lag/spectral window, $k_M(\tau)$, and the lag/spectral window parameter estimator, $\hat{M}$, so that they fulfill conditions (ii), (iii) and (iv) of spectral estimates as stated in Appendix 1. In this we follow Priestley (1981) and Andrews (1991) who show that the Quadratic Spectral

[5]Except for the seed, to avoid the possibility that the random number series were exactly the same for ϵ_t than for u_t^i. Instead, the u_t^i series have been generated with the same random number generator, starting from the point where the ϵ_t series ended. This way, ϵ_t and u_t^i, i = 1,2 and 3, are independent r.v. and inference can be constructed on the distance between transformations of them.

window is optimal (see comment on selection of the appropriate lag/spectral window function in Appendix 1). We also follow Andrews (1991) in choosing an "automatic bandwidth estimator $\hat{M}$" which is a function of the data and asymptotically optimal under general conditions (see comment on the estimator for the lag/spectral window parameter, also in Appendix 1).

3.4.1 Fit Test

Table 3.1 reports the finite sample behavior of the fit test, when the null is that the model evaluated follows a VAR(1) with same parameter values as the actual DGP. The numbers displayed in Table 3.1 are the percentage rejection across 1000 Monte Carlo replicaitons of the null hypothesis of zero distance between the model and the observed data for the tests $fit(\omega;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$. Under "Model i" (i = 1,2 and 3) we compute the test statistic comparing the spectral density matrix of $\{y_t\}_{t=1}^T$ to the one of $\{x_t^i\}_{t=1}^T$. Therefore, the numbers under "Model 1" measure the *size* of the fit testand the rest measure the *power* under different alternative hypothesis,i.e. x_t^i coming from Models 2 or 3.

The performance of each model, either correctly specified (Model 1) or not (models 2 and 3), with respect to the actual data is evaluated for one single frequency (ω_1 is the frequency associated with cycles of periodicity 8 years, and ω_2 with those of periodicity 2 years) and for the inclusive business cycle set of frequencies ($[\omega_1, \omega_2]$). The critical values used with $fit(\omega;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$ are the critical values of a $\chi^2_{(N^2)}$ and a $\chi^2_{(N^2L)}$ distribution, respectively, which correspond to the theoretical size indicated (either 5% or 10%). N=2 in all cases. The quantity L depends on the lag/spectral window parameter $\hat{M}$ we use in the estimation of the spectral density matrix as explained in Appendix 1. $\hat{M}$ depends on the length of the series, T, as well as on the parametric DGP z_t is supposed to follow. This guarantees that spectral estimates at different frequencies are independent (see comment on property (c) at Appendix 1). In particular, we estimate the crosspectra at frequencies distant $\frac{\pi}{\hat{M}}$ to each other. For a sample size T of 100, 200 and 500 observations, the estimation procedure followed here yields L=4, 4 and 5 frequency points, respectively.

Table 3.1 shows that the *size* of the test (panel "Model 1") is found smaller than

its theoretical value even for sample size of 100 observations. This indicates that the empirical distribution of the test statistic is skewed relative to the theoretical one, more concentrated around values closer to zero. The *size* of the aggregated version of the test, $fit([\omega_1,\omega_2];\gamma)$, is poor for sample size T=100. The reason is simply that we are aggregating the small sample biases of the spectral estimates each single-frequency test statistic carries[6]. Howevever, this effect disappears fast as the sample size increases.

One feature Table 3.1 shows is the low percentage rejection of models 2 and 3 at the frequency associated roughly to cycles of 2 years length i.e. ω_2. This should not be interpreted as a low power of the test but as the ability of the test to capture the fact that although two models may look different in the time domain, they may generate similar dynamics for a particular frequency range. In this case, with respect to Model 2, the two VAR models have very similar spectra except at the lower frequencies. For the particular DGP considered, ω_2 is outside these lower frequencies and therefore discrimination between different VAR models is difficult (low rejection frequency). A White Noise process (Model 3) has a flat spectral function which can have at a particular frequency the same power density as a VAR, with similar values as well in the neighborhood of that frequency. These two features can be more easily appreciated looking at Figure 3.1. Figure 3.1 shows the average, across the 1000 replications of the Monte Carlo experiments, of the $fit(\omega;\gamma)$ test-statistics for different sample sizes and for the whole frequency range. All values are transformed in logs. The horizontal line corresponds to the critical value. The solid line under the critical value is the test-statistic for Model 1. The discontinuous and starred lines correspond to Models 2 and 3, respectively. It can be seen that the distance between Models 1 and 2 (the other VAR, dashed line) is not significant apart from the very low frequencies, and that the distance between the DGP and Model 3 (starred line) is not significant in the vicinity of frequency ω_2, for the reasons explained above.

The aggregate versions of the test, $fit([\omega_1,\omega_2];\gamma)$, performs in general worse than the one-frequency version when evaluated at ω_1 but substantially better than when evaluated at ω_2. Because of the possibility of a spurious coincidence of the spectra of two different models around a particular frequency (as noticed around ω_2 when

[6]This small sample bias has also been found in existing Monte Carlo studies of other kinds of estimation procedures, such as GMM (see Andersen and Sorensen (1995)).

evaluating Model 3), the *fit* test appears clearly more powerful when used to examine the performance of a model in a frequency band rather than when used to assess the fit of a model at ω_2.

It can be seen that is the *fit* test manages to correctly rank the models according to the actual DGP: the rejection frequency of Model 1 (equal to the actual DGP) is lower than that of Model 2 (a VAR(1) as Model 1 but with different parameter structure), which in turn is rejected as similar to the actual DGP a lower number of times than Model 3 (bivariate WN) in almost all cases. The particular value of the test statistic (alternatively the associated p-value) can therefore be considered as a *ranking device*: the further away is that value from the fixed critical value (or the significance level) the worse is the performance of the Model with respect to the particular actual data set and for the selected frequencies. Given that we may be interested in evaluating models that almost surely differ from the true DGP, as the majority of multivariate dynamic models in many fields of economics, it is of great interest to evaluate how "poor" is their performance with respect to alternative models aimed at explaining the same relationships observed in the actual data. It is in these cases where the *fit* test can be used most usefully as a ranking device, with respect to an arbitrary lower bound B (which we consider to be the best fit possible). In particular, B would be the critical value of the non-central χ^2 distribution we referred to in section 3.2, once we have fixed the arbitrary minimum distance we expect of the evaluated model to the actual data, C.

We have also performed some sensitivity analysis on the choice of parameter values for the DGP. If one is interested in capturing the cross-variable relationships observed in the actual data (e.g. interdependencies between the cyclical properties of the GDP of two countries) it is important to check whether our test is able to discriminate between models which give different degrees of interdependence between variables. In the VAR framework, a higher degree of cross-variable dependence can be translated into higher off-diagonal coefficients in the Φ matrix ("spillover" coefficients). Table 3.2 displays the corresponding Monte Carlo rejection frequencies when the actual data DGP has a different VAR structure:

I) No spillovers: $\Phi^y = \Phi^2 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.6 \end{pmatrix}$

II) Larger spillovers: $\Phi^{\mathbf{y}} = \begin{pmatrix} 0.7 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$

Figures 3.2 and 3.3 represent the log of the average $fit(\omega;\gamma)$ test statistic across the 1000 Monte Carlo replications for Case I and II, respectively.

For Case I the two series in both bivariate vectors $\mathbf{y}_t$ and $\mathbf{x}_t^1$ follow independent AR(1) processes. Then, the theoretical power spectrum of the actual data series is more concentrated at the lower frequencies (hence more distinguishable from other models, yielding higher power for the test) while keeping the difficulty to discriminate a VAR from a WN at certain frequencies close to ω_2. However, as the DGP displays smaller interdependences, $\hat{M}$ becomes smaller (only low order covariance estimates needed to be included in $\hat{f}(\omega;\gamma)$ to capture the DGP dynamics). Given that the number of independent frequency point estimates decreases the smaller $\hat{M}$ is, we obtain estimates of the spectral density matrices (and hence of the test statistics) for a smaller number of frequencies than when using other DGPs. Therefore comparisons between the empirical rejection frequencies obtained with other DGPs have to be made with care.

Under Case II interdependence between the series in both $\mathbf{y}_t$ and $\mathbf{x}_t^1$ vectors increases. The spectral power of the the true DGP becomes less concentrated in the lower frequencies the higher the interdependence because it increases the relative magnitude of lagged covariances with respect to the contemporaneous (spectrum or crosspectrum at frequency zero). As a result, the true spectra and crosspectra are less distinguishable from the WN Model (equal power spectra and crosspectra at every frequency). In general, the Monte Carlo experiments for this case show lower size and power than under Case I, especially for the shorter sample sizes considered, but the other features we have described are unchanged. However, there is an important effect on the automatic bandwidth estimator: $\hat{M}$ increases to capture this higher interdependence in the DGP (to include higher order covariances in $\hat{f}(\omega;\gamma)$) and hence the number of frequency point estimates increases. The difference in the number of frequencies estimated with respect to Case I suggests again that care should be taken in comparing the numbers in Table 3.2 with those in Table 3.1. This confirms the influence of the characteristics of spectral density estimates (e.g. the effect of $\hat{M}$) in the small sample performance of the *fit* test relative to other aspects of the model evaluation methodology we proposed. Andersen and Sorensen (1995) reach similar conclusions. Also Christiano and

den Haan (1995) who study the sensitivity of the small sample bias in GMM estimation of Business Cycle models to estimation tools used in this paper too, such as the choice of the lag/spectral window and of the bandwidth parameter M.

3.4.2 Comparison test

The comparison test has been introduced as a formal device to assess the difference between the dynamic properties of multivariate time series generated by alternative model specifications.

Table 3.3 displays the percentage rejection of the null of no difference between the spectral density matrix of the multivariate series of interest when they are simulated from model specification (m_i, γ_i), $\{x_t^i\}_{t=1}^T$, and from (m_j, γ_j), $\{x^j\}_{t=1}^T$, for i,j = 1,2,3. When i = j we report the *size* of the test and when i $\neq$ j, its *power*. The first two blocks of Table 3.3 (Case11 and Case33) compute the size of the test when the models are equal to the DGP of the actual data (both are Model 1) and when they are equal but both misspecified (both are Model 3). Note that the null hypothesis of the comparison test is that both models have the same spectral density matrix for the multivariate vector of interest, not that this spectral density matrix has a particular array of values. Hence, the test can be applied equally to correctly specified or to misspecified models. The last two blocks compute the power of the test when the two models are VAR(1) but with different parameters (Case12: one is Model 1 and the other Model 2, and Case13: one is Model 1 and the other a bivariate WN -Model 3-).

Because the asymptotic distribution of the comparison test is the same as that of the fit test, the critical values used are the same as in subsection 3.4.1.

As in the case of the fit test, the *size* becomes smaller the larger the sample size. Note that the empirical size of the test in Case33 (both models WN) is smaller than in Case11 (both having the same VAR structure). Probably, the main reason for this difference is that the dynamic structure assumed for z_t in the test is more complicated under Case11 and therefore bound to induce higher small sample bias than when the DGP of the simulated series is a WN (Case33). Further research in terms of DGP parameter sensitivity would clarify this point.

Here again, it is very useful and intuitive to represent graphically the comparison test-statistic under different specifications. Similarly to Figure 3.1, Figure 3.4 shows the average value for the $comp(\omega;\gamma)$ test-statistic across the 1000 Monte Carlo replications, and for different sample sizes. As before, all values are transformed in logs. The horizontal lines correspond to the 90% and 95% critical value for the one-frequency test. The solid (starred) line under the critical value is the test statistic for the case in which Model i and and Model j are generated from the same DGP, both are Model 1 (Model 3). The discontinuous (dashdotted) line corresponds to Case12 (Case13). We can see that the test has difficulties to discriminate between a WN and a VAR in the frequencies immediately higher than $\frac{\pi}{2}$; the *comp* test-statistic lies under the critical value line. This is the same type of effects observed in figure 3.1: a VAR(1) such as Model 1 can imply at some frequency the same power density as the flat spectrum of a WN (Model 3). As we found for the *fit* test, the performance of the *comp* test depends substantially on the characteristics of the spectral estimates.

Also as in the *fit* test case, it is remarkable how informative the value of the rejection frequency is about how different the model specifications compared are. The *comp* test is found for all sample sizes more powerful rejecting a WN model as equal to a VAR(1) (Case13) than two different VAR(1) as equal to each other (Case12).

3.5 An example

In this section we apply the *fit* and *comp* tests to versions of a standard International Real Business Cycle (IRBC) model to assess how they reproduce the interdependencies observed between the US and European business cycles over the period 1970Q1-1993Q3. We measure these interdependencies estimating the bivariate spectral density matrix of the US and European real GDPs at business cycle frequencies (those associated to cycles 8 to 2 years long). We have used the Quadratic Spectral density window and the Andrews' optimal bandwidth parameter. Figure 3.5 plots the linearly detrended logs of the series (quarterly seasonally adjusted real GDP from OECD Main Economic Indicators). The information contained in the estimated spectral density matrix is rearranged in Figure 3.6, which shows the individual spectra of the two GDPs, the *phase* (whether there exists a lead or lag between the two series) and the *coherence*

(the equivalent to the correlation in the frequency domain, also varying between 0 and 1) for all frequencies. The business cycle frequency range is indicated with vertical bars in the phase and coherence plots. All statistics are plotted with their corresponding asymptotic confidence intervals. In the case of the coherence, the lower (upper) horizontal line is the 95% (99%) critical value of the test H_0: coherence=0.

The European GDP is found substantially more volatile than the US one (higher values of the spectrum). The short sample size (95 observations) causes an imprecise estimate in both cases: the 95% asymptotic confidence interval bands are wide and the lower one is not distinguishable from 0 for all frequencies. A significant coherence is found between the two GDPs at all business cycle frequencies (and for most of them, even at a 99% confidence level), altough none of the GDPs clearly lead the other one. The well known "locomotive role" of the US economy with respect to the european one is not clear for this period. Figure 3.5 reveals that this is so because the two GDPs where evolving very synchronized in the 70s and it is only in the 80s when the US GDP clearly leads the European one.

3.5.1 The models

The benchmark model economy is a standard two-country two-good International Real Business Cycle model. The possible sources of economic fluctuations in the model are stochastic shocks from both the demand and the supply side of the economy, eihter country-specific or common (i.e. contemporaneous cross-country correlation between shocks). Demand disturbances take the form of exogenous government expenditure shocks while supply disturbances are identified with technology shocks. The specific mechanisms of international transmission of shocks and fluctuations allowed in the model are trade in final goods and services and spillovers in the shocks processes.

Each country specializes in the production of a single differenciated good, in the lines of Backus, Kehoe and Kydland (1993). Each country is populated by a large number of utility maximizers infinitely-lived identical agents. The representative household sells the services of capital and labor, owns all the firms and receives all the profits. The goods produced by the firms will be purchased by the household in order to be consumed or invested. There are complete financial markets within countries and free

mobility of physical and financial capital across countries. However, labor is immobile internationally. Each household in country i has preferences given by the utility function

$$U_{it} = E_t \sum_{i=0}^{\infty} \beta^i (1-\sigma)^{-1} (C_{it+1}^{\theta} L_{it+1}^{1-\theta})^{1-\sigma} \tag{3.19}$$

where C_{it} is consumption at time t, L_{it} is leisure, $0 < \theta < 1$ is the relative weight of consumption to leisure and σ is the risk aversion parameter. The endowment of time is H_t in each period, which constrains leisure to be between 0 and H_t.

There is a representative firm operating in each country that produces output with a constant returns-to-scale production function

$$Y_{it} = A_{it} K_{it}^{\alpha} (X_{it} N_{it})^{1-\alpha} \tag{3.20}$$

where K_{it} and N_{it} are capital and labor used by firms in country i and α is a parameter governing the output share of capital. X_{it} represents the state of technology at time t. Total factor productivity, A_{it}, follows the joint process

$$\begin{bmatrix} lnA_{1t} \\ lnA_{2t} \end{bmatrix} = \begin{bmatrix} \rho_{A1} & \nu_{12} \\ \nu_{21} & \rho_{A2} \end{bmatrix} \begin{bmatrix} lnA_{1t-1} \\ lnA_{2t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}, \quad \epsilon_t \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\epsilon 1}^2 & \psi \\ \psi & \sigma_{\epsilon 2}^2 \end{bmatrix} \right) \tag{3.21}$$

where ρ_{Ai} is the parameter that governs the persistence of the technology process within country h, ν_{ij} is the spillover parameter determining the speed at which changes in technology in country i are transmitted to country j, $\sigma_{\epsilon i}$ is the standard deviation of the stationary exogenous technology shocks in country i, ϵ_{it}, and ψ represents the covariance between the innovations to technology processes in both countries, i.e. a common shock to both countries total factor productivities will be characterized by a high ψ, and the higher ψ the less country-specific is the shock.

Multiple goods are introduced by assuming that Y_{it} can be either used domestically or exported

$$Y_{1t} = A_{1t} + \frac{\Pi_2}{\Pi_1} \tilde{A}_{2t}, \quad Y_{2t} = \frac{\Pi_1}{\Pi_2} \tilde{B}_{1t} + B_{2t}$$

where $\tilde{A}_{2t}$ and B_{1t} are exports and imports of country 1, and Π_i represents the size of each country, e.g. population. Let $A_{2t} = \frac{\Pi_2}{\Pi_1} \tilde{A}_{2t}$ and $B_{1t} = \frac{\Pi_1}{\Pi_2} \tilde{B}_{1t}$. Imports and domestic goods are used in the production of a final good in each country, V_{it}, according

to the following constant elasticity of substitution technology (see Armington (1969)):

$$V_{1t} = (\omega_1 A_{1t}^{1-\rho} + \omega_2 B_{1t}^{1-\rho})^{\frac{1}{1-\rho}}, \quad V_{2t} = (\omega_1 B_{2t}^{1-\rho} + \omega_2 A_{2t}^{1-\rho})^{\frac{1}{1-\rho}} \tag{3.22}$$

where $\frac{1}{\rho}$ is the elasticity of substitution between domestic and foreign goods and ω_1 and ω_2 are parameters regulating the domestic and foreign content of GNP. This is a very convenient specification because it allows for cross-hauling (a situation in which a country imports and exports goods of the same category) and permits both countries to produce same categories of goods. The relative price of imports to exports (terms of trade) is given by

$$P_{1t} = \frac{\partial V_1 / \partial B_{1t}}{\partial V_1 / \partial A_{1t}} = \frac{\omega_2 B_{1t}^{-\rho}}{\omega_1 A_{1t}^{-\rho}}$$

where $\omega_1 = (1-\text{MS})^\rho$, $\omega_2 = \text{MS}^\rho$ and MS is the import share in output. A value of MS of zero would mean that there is no trade between the economies (autarky).

Firms accumulate capital goods according to the following law of motion

$$K_{i,t+1} = (1-\delta)K_{it} + I_{it} \tag{3.23}$$

where K_{it} is the total stock of capital in country i, $0 < \delta_i < 1$ is the rate of depreciation of capital stock and I_{it} is total investment in country i.

In addition to consumers and producers, each country is endowed with a government. The government consumes domestic goods (G_{it}), taxes national output with a distortionary proportional income tax (τ_{it}) and transfers back the remaining to domestic residents (T_{it}). Alternatively, the government can issue debt that will be repaid by increases in lump-sum taxes or decreases in transfers. The infinite horizon of this economy makes Ricardian equivalence hold: it is equivalent to finance government expenditure with taxes or with debt that will be compensated in the future with either more taxes or less transfers. The government flow budget constraint is given by

$$G_{it} + T_{it} = \tau_{it} Y_{it} \tag{3.24}$$

which has to hold on a period by period basis. Government spending is assumed to follow an autoregressive stochastic process of the form

$$\begin{bmatrix} G_{1t} \\ G_{2t} \end{bmatrix} = \begin{bmatrix} \rho_{G1} & \nu_{G,12} \\ \nu_{G,21} & \rho_{G2} \end{bmatrix} \begin{bmatrix} G_{1t-1} \\ G_{2t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{G1t} \\ \epsilon_{G2t} \end{bmatrix}, \quad \epsilon_{Gt} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\epsilon_{G1}} & \psi_{G,12} \\ \psi_{G,21} & \sigma^2_{\epsilon_{G2}} \end{bmatrix} \right) \tag{3.25}$$

where ν_G controls for the spillover effect from government spending and $\psi_G \neq 0$ means that we allow for common fiscal policy shocks. ρ_{Gi} is the persistence parameter of the government spending process and $\sigma_{\epsilon_{Gi}}$ is the standard deviation of the innovation to the government spending process in country i.

There is frictionless international trade and capital markets are complete, which implies that individuals in the two countries can achieve both consumption smoothing (intertemporal transfer of consumption) and risk pooling (transfer of consumption across states of nature). The trade balance, or net exports, in country i is then given by $NX_{it} = Y_{it} - (C_{it} + I_{it} + G_{it})$.

Finally, the aggregate resources constraint for the traded goods in the world economy is

$$\Pi_1(C_{1t} + I_{1t} + G_{1t}) + \Pi_2(C_{2t} + I_{2t} + G_{2t}) = \Pi_1 V_{1t} + \Pi_2 V_{2t} \tag{3.26}$$

The equilibrium solution of the model can be obtained by deriving the sequences for the endogenous variables of the model that maximize (3.19) subject to (3.20)-(3.26) and given the initial endowment of capital. The complexity of the economic relations described in the model (highly nonlinear) causes, as in most RBC models, that an analytical solution which derives the endogenous variables in terms of the exogenous and the parameters cannot be obtained. We follow King, Plosser and Rebelo (1988) and use an Euler equation approximation approach that log-linearizes the set of first order conditions of the model, expressed in terms of "detrended" variables (all trending variables are transformed into ratios of the permanent technology change x_{it}), around the steady state. Once the approximate solution is found, the parameters are given fixed values and solution paths for the variables of interest are simulated from the model drawing realizations from the exogenous processes' fixed distributions.

Parameter values are chosen from the IRBC literature, basically from Backus, Kehoe and Kydland (BKK) (1992)-(1993) and Ravn (1993). The representative agent's discount factor, β, is 0.9875; the coefficient of relative risk aversion, σ, is 2; the steady state share of time the household allocates to market activities is 30%. On the production side of the economy, the output share of capital α, is 0.36 and the quarterly depreciation rate, δ, is 0.025%. BKK (1992), Ravn (1993) and Reynolds (1993) have constructed several measures of aggregate Solow residuals for different measures of pro-

duction factors and selected pairs of countries. They all obtain values of persistence (ρ_A) very close to one and positive correlations (ψ) across technology shocks to different countries. In addition, both BKK (1992) and Ravn (1993) find evidence of significant spillovers (ν_{ij}), in particular from the US to other countries, although Reynolds (1993) suggests that these spillovers may be quite low. We follow BKK (1992) whose estimate for the persistence parameter, ρ_A is 0.906, for the spillover parameter ν is 0.088, for the standard deviation of the technology shock σ_ϵ is 0.00852 and for the cross-country correlation between shocks is 0.258 ($\nu_{12} = \nu_{21}$). Parameters for the government sector are taken from Baxter and King (1993), Aiyagari, Christiano and Eichenbaum (1992) and King, Plosser and Rebelo (1988). We impose government budget balance in the steady state by assuming a constant tax rate (τ) equal to a constant government spending output share (sg). We have taken a value for τ and sg of 25%, which lays in between the one suggested by King, Plosser and Rebelo (1988) of 30% and that used by Baxter and King (1993) of 20% for the case of steady state balanced budget (Aiyagari, Christiano and Eichenbaum (1992) suggest a government spending share of 17.7%). The steady state import share, MS is set to 15%, and the parameter governing the elasticity of substitution in the Armington aggregator, ρ, to 1.5. Finally, we assume that both countries have equal size, $\Pi = 1/2$.

By modifying certain key parameters which govern the international interdependencies included in the model, we derive four different model specifications:

(i) Autarky: No trade (MS=0), no spillovers ($\nu_{ij} = 0$) and uncorrelated shocks (no common shocks).

(iii) Autarky with common shocks: No trade (MS=0), no spillovers ($\nu_{ij} = 0$) but contemporaneously correlated technology shocks ($\psi \neq 0$). There are common and country-specific shocks to closed economies.

(iii) Trade: No spillovers ($\nu_{ij} = 0$) nor common shocks ($\psi = 0$) but trade in final goods and services is allowed between the two economies (MS $\neq 0$).

(iv) Full interdependence: common and country-specific shocks transmitted through trade and through spillovers in the shock processes (ν_{ij}, MS and $\psi \neq 0$).

Table 3.4 summarizes the parameter vector under each model specification.

In order to characterize the international interdependencies generated by the four

models, we perform the same multivariate spectral analysis of Figure 3.6 to the output series generated for the two countries under each model specification. Figures 3.7 and 3.8 plot the individual spectra, phases and coherences of the linearly detrended logs of the two output series simulated once under each of the four alternative model specifications. As with the actual data, we obtain imprecise estimates of the output spectra (wide bands) under all specifications. However, it is clear that model spectra are of similar size for the two countries, contrary to what is observed in actual data (European GDP far more volatile than US one). All but the "Autarky with common shock" specification predict a lead of the GDP of one country over the other one, which we saw is not the case in actual data. However, the four models are able to capture the sigificant coherence (at 95% confidence level) between GDPs at business cycle frequencies, and predict values similar to the actual coherence. It is the "full interdependence" model the one displaying the higher coherence (significantly different from zero even at a 99% confidence level).

3.5.2 Assessment of the models

Assessment focusses on how well each model reproduces the dynamic relationship between both GDPs at business cycle frequencies. We have simulated the model 100 times (H=100) and estimated at each simulation the measure of distance $\hat{D}_h(\omega;\gamma)$ between the actual and simulated spectral density for both countries' GDPs, where γ is the corresponding column in Table 3.4. The fit of the model is computed at each frequency based on the average estimated distance across simulations, $\hat{D}(\omega;\gamma)$.

We first apply the *fit* test to all four model specifications. The average fit across business cycle frequencies for each model is displayed in the diagonal of Table 3.5. The 90% and 95% critical values for the test statistics can be found in the bottom part of the table. They correspond to a χ^2 distribution with degrees of freedom 2^2*7, sine there are 2 variables and the Andrews' optimal bandwidth estimator yields 7 frequencies in the business cycle frequency range.

The standard IRBC two-country model with multiple goods has the best fit when common shocks are the only mechanism by which economies may move together (there are no spillovers across national shocks nor trade in final goods), but still is clearly

rejected as the US-Europe DGP with respect to the comovements of the two countries' real GDPs at business cycle frequencies. Trade in final goods is clearly not the main mechanism by which fluctuations are transmitted among the US and European economies: our methodology assigns to the "trade only" model the worse fit. Figure 3.9 plots the *fit* test statistic for each model specification and for all frequencies. The horizontal lines are the 90% and 95% critical values of the one-frequency test (from a χ^2_4 distribution). It can be seen that all models are found similarly distant to the actual data and that it is the model using common shocks as the only explanation of international comovements the one reaching values of the fit test which get closer to the critical values at business cycle frequencies.

The off-diagonal figures in Table 3.9 are the $comp([\omega_1, \omega_2])$ test statistic applied to compare the performance (failure) of the four IRBC models two by two, for the business cycle frequency range, too. Their frequency by frequency values are plotted with the critical values in Figure 3.10. Any model is clearly rejected as equal to any other, with test statistics indicating in all cases that they are further to other models than to the actual data. Aggregating across business cycle frequencies suggests that the closer two models are the "autarky" and the "full interdependence" ones, but Figure 3.10 shows that the "full interdependence" model is closer to that including only "common shocks" at the lower frequencies. The more distant models are, logically, those with the best and worse fit: "common shocks" to "trade only" models.

This example borrowed from the IRBC literature allows us to illustrate how the *fit* test statistic proposed in this chapter assesses the fit of a dynamic general equilibrium model and is able to produce a ranking of competing models. In particular, it has evidenced the importance of the existence of common shocks to explain the significant comovement observed between the US and European economies in 1970Q1-1993Q3 (high coherence between their real GDPs at business cycle frequencies). A simple comparison of the spectral density properties of the four models we have studied was able to conclude that all of them where quite insatisfactorily reproducing those of the actual data, but was not sufficient to discriminate across model specifications. Our model evaluation methodology confirms statistically the rejection of all models (fit statistics greater than their asymptotic critical value) and manages to identify that it

is the one including common shocks as the only source of international comovements the one having best fit.

3.6 Summary and Conclusions

This chapter develops a general formal framework for assessing the fit of multivariate dynamic models whose solution is approximated through simulation. The procedure is based on multivariate spectral techniques which are especially suitable for models that focus on a particular frequency range, such as business cycle models. The test we propose evaluates the distance between the spectral density matrices of the actual data and of data simulated from a model. Important features of the test are that it treats the sample of observed data and the simulated series symmetrically, and that it formally takes into account the fact that model series are simulated with an approximation error.

The necessary asymptotic theory is derived to test how well a simulated model reproduces the dynamic properties of actual data (*fit* test). Another asymptotic test is derived to compare the performance of alternative model specifications with respect to the multivariate vector of interest (*comparison* test). Monte Carlo evidence is provided showing the finite sample behavior of the tests. We find that both tests are very powerful against different alternatives even with small sample sizes. Sensitivity analysis shows the robustness of this result to alternative DGP specifications. However, the empirical distribution of the tests seems to be more concentrated around zero than the theoretical asymptotic distribution (which implies lower size in spite of the high power). This raises the issue of whether for the particular parametric models and sample sizes we have chosen, the measure of distance defined is still closer to a χ^2 distribution than to its limiting Normal distribution. Further research in terms of sensitivity of the methodology to the DGP structure and the sample size would clarify further this point.

Confirming other studies, we also find that the small sample properties of our tests depend on the small sample characteristics of spectral estimators, in particular on the bandwidth parameter.

We have illustrated the use of the *fit* and *comparison* test statistics to assess a

model by evaluating to which extent different versions of a two-country two-good International Real Business Cycle model can reproduce the interdependencies observed between the US and European real GDP at business cycle frequencies. Our model evaluation methodology confirms statistically the rejection of all models and manages to produce a clear ranking of competing models according to their fit, which could not be done in our case with simple inspection of the spectral densities of the actual and simulated data. The existence of common shocks is found important to explain the significant comovement observed between the US and European economies in 1970Q1-1993Q3 (high coherence between their real GDPs at business cycle frequencies).

3.7 Appendix 1: Asymptotic properties of spectral estimators.

Here we report a standard general theorem which determines the asymptotic distribution of spectral estimators and we discuss its properties. For more complete references see, for example, ch.IV and V in Hannan (1970), ch.6 and 9 in Priestley (1981), or Andrews (1991).

For any multivariate random process z_t satisfying:

(i) z_t is a zero mean multivariate general linear process

$$z_t = \sum_{\tau=-\infty}^{+\infty} A_\tau \epsilon_{t-\tau}, \tag{3.27}$$

where $\sum_{\tau=-\infty}^{+\infty} |A_\tau| < \infty$, and ϵ_t are iid processes with finite 4th order moments, and $\sum_{\tau=-\infty}^{+\infty} ||\sigma_{ij}(\tau)|| < \infty$, where $\sigma_{ij}(\tau)$ is the covariance for lag τ between z_{it} and z_{jt}, i,j = 1,..., N,

(ii) the lag window function $k_M(\tau)$ is a real valued continuous uniformly bounded function such that:

- $k_M(0) = 1$,
- $k_M(\tau) = k_M(-\tau), \forall \tau$,

- $\int_{\tau=-\infty}^{+\infty} k_M^2(\tau)d\tau < \infty$ and,
- at each ω, the corresponding spectral window function is defined by

$$W_M(\omega) = \frac{1}{2\pi}\int_{\tau=-\infty}^{+\infty} k_M^2(\tau)e^{-i\tau\omega}d\tau \geq 0, \forall\omega. \tag{3.28}$$

(iii) the *lag/spectral window parameter* M is such that $M \to \infty$ as $T \to \infty$, and

(iv) M small relative to T, so that $M/\sqrt{T} \to 0$ as $M, T \to \infty$, i.e. $\mathrm{M} = \mathrm{o}(T^{1/2})$,

the spectral estimator:

$$\hat{f}_{ij}(\omega) = \frac{1}{2\pi}\sum_{\tau=-T+1}^{T-1} k_M(\tau)\hat{\sigma}_{ij}(\tau)e^{-i\omega\tau} \tag{3.29}$$

where $\hat{\sigma}_{ij}(\tau)$ is the covariance estimate for lag τ between z_i and z_j,

has the following properties:

(a)

$$\sqrt{\frac{T}{M}}\left(\hat{f}_{ij}(\omega) - E[\hat{f}_{ij}(\omega)]\right) \sim \text{Complex Normal}, \tag{3.30}$$

(b)

$$\lim_{T\to\infty} E[\hat{f}_{ij}(\omega)] = f_{ij}(\omega) \tag{3.31}$$

(c)

$$\lim_{T\to\infty}\left(\frac{T}{M}cov\left[\hat{f}_{ij}(\omega_1), \hat{f}_{kl}(\omega_2)\right]\right) = 0, \quad \omega_1 \neq \pm\omega_2, \tag{3.32}$$

(d)

$$\begin{aligned}\lim_{T\to\infty}\left[\frac{T}{M}cov\left[\hat{f}_{ij}(\omega), \hat{f}_{kl}(\omega)\right]\right) &= \\ = \left(\int_{\tau=-\infty}^{+\infty} k_M^2(\tau)d\tau\right) f_{ik}(\omega)\overline{f_{jl}(\omega)}, \quad \omega \neq 0, \pm\pi & \\ = 2\left(\int_{\tau=-\infty}^{+\infty} k_M^2(\tau)d\tau\right)\left(f_{ik}(\omega)\overline{f_{jl}(\omega)} + f_{il}(\omega)\overline{f_{jk}(\omega)}\right), \quad \omega = 0, \pm\pi & \end{aligned} \tag{3.33}$$

where $\overline{f(\omega)}$ stands for the complex conjugate of $f(\omega)$.

Property (a) is the key to derive the **asymptotic distribution** of our measure of distance defined in (3.8). A standard result in spectral estimation is that the asymptotic distribution of $\nu\frac{\hat{f}_{ij}(\omega)}{f_{ij}(\omega)}$ can be approximated by χ^2_ν. ν is a constant called *equivalent degrees of freedom* of the spectral estimate and is defined as $\nu = \frac{2T}{M\int_{\tau=-\infty}^{+\infty} k_M^2(\tau)d\tau}$. The value of ν for each lag/spectral window function is tabulated (see Priestley (1981)), e.g. $\frac{5T}{3M}$ for the case of the Quadratic Spectral (QS) window, which is the one we will use in this paper.

However, for spectral estimates satisfying condition (iv), i.e. $M/T \to 0$ as $M, T \to \infty$, $\nu \to \infty$ and therefore the χ^2_ν distribution tends to a Normal distribution (see Priestley(1981), ch.6.4). Hence, $\hat{f}_{ij}(\omega)$ has an asymptotic Normal distribution, as in property (a).

Property (b) states the **asymptotic unbiasedness** of spectral estimates. This property holds because of assumption (iii): whichever the lag/spectral window function used, the asymptotic unbiasedness of spectral estimates is guaranteed as long as $M \to \infty$, or if M is a function of T (as it is generally the case, e.g. QS window) as long as $M \to \infty$ as $T \to \infty$.

Property (c) indicates the **independence of spectral estimates at different frequencies**. This property holds in general only when the separation between frequencies is greater than the bandwidth of the spectral window. In the case of the QS window, this requirement is fulfilled estimating the spectrum at frequencies distant $\frac{\pi}{M}$ to each other, which is the criterion we use in the Monte Carlo simulations of Section 3.4.

Property (d) derives the **asymptotic variance-covariance structure** of multivariate spectral estimates. The use of a spectral estimator with a lag/spectral window function as defined in assumption (ii) is introduced to overcome the asymptotic inconsistency of periodogram estimates. A lag/spectral window weights the sample co-

variances in the spectral estimator in (3.29) so as to reduce the contribution of distant lags/frequencies (and omits the lags/frequencies distant more than what the parameter M indicates), and thus the variance of spectral estimates is reduced. Assumption (iv), together with (iii), controls for the asymptotic properties of the lag/spectral window so that the variances and covariances of spectral estimates $\rightarrow 0$ as $T \rightarrow \infty$. These assumptions guarantee that property (d), together with (b), imply **consistency** of spectral estimates.

Regarding the condition required in assumption (iv), i.e. $M = o(T^{1/2})$, Andrews (1991) shows that optimal growth rates of M are typically less than $T^{1/2}$. He devices an automatic estimator for the **lag/spectral window parameter M**, $\hat{M}$, as a function of the sample size T and of the parametric DGP the z_t vector is supposed to follow, that is optimal under general conditions and is e.g. $O(T^{1/3})$ for the Bartlett window and $O(T^{1/5})$ for the Parzen, Tukey-Hanning and QS windows. His automatic estimator is used in the Monte Carlo experiments of section 3.4.

The **selection of the appropriate lag/spectral window function** is a controversial issue in the frequency domain literature. In this paper we want a lag/spectral window function that satisfies assumption (ii). Those conditions are satisfied by almost all standard "scale windows" $k(\frac{\tau}{M})$ used in practice, e.g. Bartlett, Parzen, Quadratic Spectral(QS) andTukey-Hanning windows.

Priestley (1981) shows that the best performing lag/spectral window, in terms of minimizing the relative Mean Square Error $(= \frac{variance^2(\omega)+bias^2(\omega)}{f^2(\omega)})$, is the Quadratic Spectral window. Andrews (1991) finds that, even when allowing for conditional heteroskedasticity and autocorrelation in the data process, the QS window is the best under a similar criterion. Therefore we use the QS window function in our simulation exercises: its functional form is (see Andrews(1991)):

$$k_{QS}(\frac{\tau}{M}) = \frac{25M^2}{12\pi^2\tau^2}\Big(\frac{sin(\frac{6\pi\tau}{5M})}{\frac{6\pi\tau}{5M}} - cos(\frac{6\pi\tau}{5M})\Big), \ \forall\tau. \tag{3.34}$$

3.8 Appendix 2: On the methodological assumptions of the *fit* and *comp* tests.

The FIT test

Typically, the econometric approach to testing models consists on testing their validity as a reduction of the DGP. This amounts to testing the null hypothesis H_0: model = DGP. As stressed elsewhere in this dissertation, such H_0 has no sense when evaluating calibrated models. Calibrators are not interested in the overall validity of the model, they are interested in whether a certain property of the model matches that of the actual data.

Then, one can still follow the traditional econometric approach and test whether a certain property of the model equals that of the actual data, e.g. the spectral density matrix of a particular set of variables, through testing H_0: sp(model) = sp(DGP). In that case, one can derive the theoretical sp(model) implied by the model from the VAR representation of the equilibrium solution for the variables of interest. Such a representation can be obtained for most models once an approximation method to obtain the equilibrium solution has been chosen, so that the dynamic properties summarized by the VAR are the ones of the model solution only under the assumption that the approximation method is exact. Then, what has to be compared to the sample estimate of sp(DGP) using the actual data is the (assumed to be exactly known) theoretical spectral density matrix of the model which is a constant matrix at each frequency. This is exactly the approach of Diebold, Ohanian and Berkowitz (1995): the only uncertainty considered to evaluate the fit of a model and to obtain afterwards optimal parameter estimates is sampling error in observed data.

Instead, the *fit* test presented in this Chapter is built under the assumption that the approximation method is not exact. Model solution is just an approximation and therefore we are not interested in obtaining its implicit theoretical spectral density matrix as fixed and true. We take the simulated series as coming from an unknown DGP of which the VAR describing the equilibrium solution is just a trustable approximation (a huge amount of effort has been devoted in the last years in the literature of dynamic stochastic general equilibrium models to how to find a trustable approximation

to the equilibrium solution to those models when its exact form cannot be obtained analytically, i.e. is "unknown"). Hence, the departure point of the *fit* test is, on the one hand, not a very well specified parametric conditional density of the model vector series but the series generated by the model solution method themselves, knowing that they can inform us about the dynamic properties of the model but with an error (the approximation error). On the other hand we have the sample of actual series.

Having those two sets of data, our intention is not to evaluate the overall validity of the model but only one specific feature: its spectral density matrix at certain frequency/frequencies. Hence, since we do not want to make assumptions on the distribution of model series (because of the approximation error), we can make use of the standard theory of spectral density estimators and build a statistic whose distribution (asymptotic) can be derived irrespective of the likelihood of both model and observed series, imposing only basic regularity conditions on both sets of data as explained in Appendix 1.

We consistently estimate the spectral density matrix both for model series and for actual data and test whether the difference is zero using the distribution of the spectral density estimator rather than that of model series. In particular, we have estimated the spectral density of the extended vector which includes model series and actual series $\hat{f}$, and then tested $H_0 : f^x - f^y = 0$ using its estimated counterpart, i.e. the difference between the submatrices of $\hat{f}$ corresponding to only model series $\hat{f}^x$ and only to actual data $\hat{f}^y$.

The COMPARISON test

For similar reasons why the *fit* test cannot be included into the traditional way of testing models, the *comp* test cannot be included into the literature of testing non-nested models. Testing non-nested hypotheses amounts to testing whether the true probability density of the observed data conditional on the exogenous variables belongs to the family of conditional probability densities of H_1 or to that of H_2. In the case of comparing two non-nested models, M1 and M2, they test H_1: M1 = DGP vs. H_2: M2 = DGP. The starting point is one well-defined likelihood for each model. The

non-nested models test statistic (say, the Cox test statistic or the Wald Encompassing Test statistic, which are equivalent when the statistic of interest in the Encompassing approach is the Likelihood Ratio test statistic, as shown in Monfardini (1995)) is built under the crucial assumption that one of the two models, say M1, is a valid reduction of the DGP so that under the hypothesis H_1: M1=DGP the Cox test statistic has a limiting χ^2 distribution.

Instead, as explained above, in the new methodology proposed in this Chapter there is no attempt to characterize the two alternative models with two parametric conditional densities whose parameters can be estimated by Pseudo Maximum Likelihood. Models are not compared using the likelihood of the vector of variables of interest. Instead, the only thing that is compared are the estimates of the spectral density matrices of the series simulated by each of the competing models, $\hat{f}^1$ and $\hat{f}^2$, whose limiting distribution is used as in the *fit* test to derive the limiting distribution of the test statistic *comp*. The null hypothesis of interest is $H_0 : f^1 - f^2 = 0$ vs. $H_1 : f^1 - f^2 \neq 0$.

There is no claim that any of the models is a valid reduction of the DGP and their spectral density matrices are compared taking into account that the series are simulated by both models with an approximation error (so that we do not compare the theoretical spectral density matrices implied by both models but their estimates and use the limiting distribution of the spectral denstiy matrix estimator).

Table 3.1: Monte Carlo on the FIT TEST

	Model 1 (size)			Model 2 (power)			Model 3 (power)		
	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$
Theoretical size: 5%									
T=100	3.1%	3.6%	13.2%	10.8%	3%	19.1%	81.1%	6.4%	87.7%
T=200	2.5%	2.2%	4.3%	16%	1.8%	16.8%	99.5%	9.8%	99.5%
T=500	1.7%	1.7%	2.5%	45.6%	3.1%	34.2%	100%	34.1%	100%
Theoretical size: 10%									
T=100	13.1%	12.3%	24.4%	25.6%	11.2%	34.5%	94%	18.2%	94.3%
T=200	8.1%	6.9%	10.8%	30.1%	6.8%	27.5%	100%	28.6%	99.8%
T=500	4.9%	4.5%	6%	62.1%	6.9%	48.7%	100%	52.8%	100%

Notes: Actual data DGP is Model 1, a bivariate VAR(1) (see description in the text).

ω_1 (ω_2) is the frequency associated to cycles 8 years (2 years) long. $[\omega_1,\omega_2]$ aggregates the test statistics for all frequencies associated to cycles between 8 to 2 years long (Business Cycle frequencies).

The Monte Carlo variance for these rejection frequency estimates is MCvar = $\sqrt{\frac{\alpha(1-\alpha)}{NREPL}}$, where α is the theoretical size and NREPL (=1000) the number of replications of the Monte Carlo experiment, i.e. MCvar = $\sqrt{\frac{0.05(1-0.05)}{1000}} = 0.69\%$ for the first panel and MCvar = $\sqrt{\frac{0.1(1-0.1)}{1000}} = 0.95\%$ for the second one.

Table 3.2: **Sensitivity of the FIT TEST to the parameter structure**

CaseI: No spillovers

	Model 2 (size)			Model 1 (power)			Model 3 (power)		
	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$
Theoretical size: 5%									
T=100	17.2%	11.8%	28.1%	49%	9.9%	51.6%	99.7%	29.2%	99.7%
T=200	8.3%	5.5%	9.5%	59.8%	6.9%	50.9%	100%	49%	100%
T=500	3.7%	2.6%	4.6%	91.1%	5.2%	78.6%	100%	37.5%	100%
Theoretical size: 10%									
T=100	29.8%	23.5%	41.7%	64.9%	23.7%	63.2%	100%	45.5%	99.8%
T=200	16.4%	12.4%	17.1%	72.2%	15%	63.8%	100%	64.8%	100%
T=500	9.2%	6.7%	8.6%	95.8%	10.2%	86.7%	100%	54.6%	100%

CaseII: Larger spillovers

	actual DGP (size)			Model 2 (power)			Model 3 (power)		
	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$
Theoretical size: 5%									
T=100	0.9%	0.7%	4.6%	14.4%	2.2%	30.3%	42.5%	3.6%	82.5%
T=200	0.7%	0.5%	1.7%	43.3%	1.8%	50.1%	85.4%	3.3%	99.3%
T=500	1.5%	1.1%	1%	96.3%	7.8%	97.3%	100%	21.3%	100%
Theoretical size: 10%									
T=100	5.5%	4.9%	12.2%	35.2%	8.5%	50.5%	71.3%	13.3%	93.1%
T=200	4.3%	2.7%	5%	67.4%	8.2%	65.1%	99.5%	12.8%	99.8%
T=500	3.6%	2.7%	2.4%	98.9%	17.2%	98.8%	100%	42.1%	100%

Notes: Actual data DGP is Model 2 in CaseI (No spillover: $\Phi^{\mathbf{y}} = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.6 \end{pmatrix}$) and a VAR(1) with parameter matrix $\Phi^{\mathbf{y}} = \begin{pmatrix} 0.7 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$ in CaseII (same DGP was used to generate model series in the cases of the column called "Actual DGP").

ω_1 (ω_2) is the frequency associated to cycles 8 years (2 years) long. $[\omega_1,\omega_2]$ aggregates the test statistics for all frequencies associated to cycles between 8 to 2 years long (Business Cycle frequencies).

The Monte Carlo variance for these rejection frequency estimates is MCvar = $\sqrt{\frac{\alpha(1-\alpha)}{NREPL}}$, where α is the theoretical size and NREPL (=1000) the number of replications of the Monte Carlo experiment, i.e. MCvar = $\sqrt{\frac{0.05(1-0.05)}{1000}} = 0.69\%$ for the first panel and MCvar = $\sqrt{\frac{0.1(1-0.1)}{1000}} = 0.95\%$ for the second one.

Table 3.3: **Monte Carlo on the COMPARISON TEST, H_0: $x^{m_i} = x^{m_j}$**

	Case11 (size)			Case33 (size, missp.)		
	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$
Theoretical size: 5%						
T=100	4.6%	3.1%	11.6%	3.9%	4.4%	11.2%
T=200	2.6%	1.8%	3.4%	1.7%	2.2%	4.6%
T=500	2.1%	1.9%	2.1%	1.2%	1.2%	2.5%
Theoretical size: 10%						
T=100	13.8%	11%	23.2%	11.8%	12.9%	22.1%
T=200	7.7%	7.7%	9.3%	6.6%	7.6%	9.3%
T=500	4.5%	5.5%	5.1%	3.7%	5.1%	4.6%
	Case12 (power)			Case13 (power)		
	ω_1	ω_2	$[\omega_1,\omega_2]$	ω_1	ω_2	$[\omega_1,\omega_2]$
Theoretical size: 5%						
T=100	10.4%	2.6%	18.6%	79.4%	6.8%	88%
T=200	14.1%	3%	14.9%	99%	11.8%	99.1%
T=500	46.2%	2.5%	33%	100%	35.7%	100%
Theoretical size: 10%						
T=100	25.5%	10%	33.4%	94.3%	18.7%	94.7%
T=200	29.2%	7.9%	26.8%	100%	27.5%	99.9%
T=500	62.5%	6.3%	46.5%	100%	52.9%	100%

Notes: Case ij indicates that we are testing the null that Model i has the same spectral density matrix than Model j, i,j=1,2 and 3. Hence, Case11 and Case33 compute the size of the test under either correct specification of the two models (Case11: both are equal to the actual data DGP, Model 1) or misspecification (Case33: the two models compared are equal, generated with Model 3, but different to the actual data DGP, Model 1). Case12 and 13 compute the power of the test for two departures from the null.

ω_1 (ω_2) is the frequency associated to cycles 8 years (2 years) long. $[\omega_1,\omega_2]$ aggregates the test statistics for all frequencies associated to cycles 8 to 2 years long (Business Cycle frequencies).

The Monte Carlo variance for these rejection frequency estimates is MCvar = $\sqrt{\frac{\alpha(1-\alpha)}{NREPL}}$, where α is the theoretical size and NREPL (=1000) the number of replications of the Monte Carlo experiment, i.e. MCvar = $\sqrt{\frac{0.05(1-0.05)}{1000}}$ = 0.69% for the 5% size case and MCvar = $\sqrt{\frac{0.1(1-0.1)}{1000}}$ = 0.95% for the 10% size case.

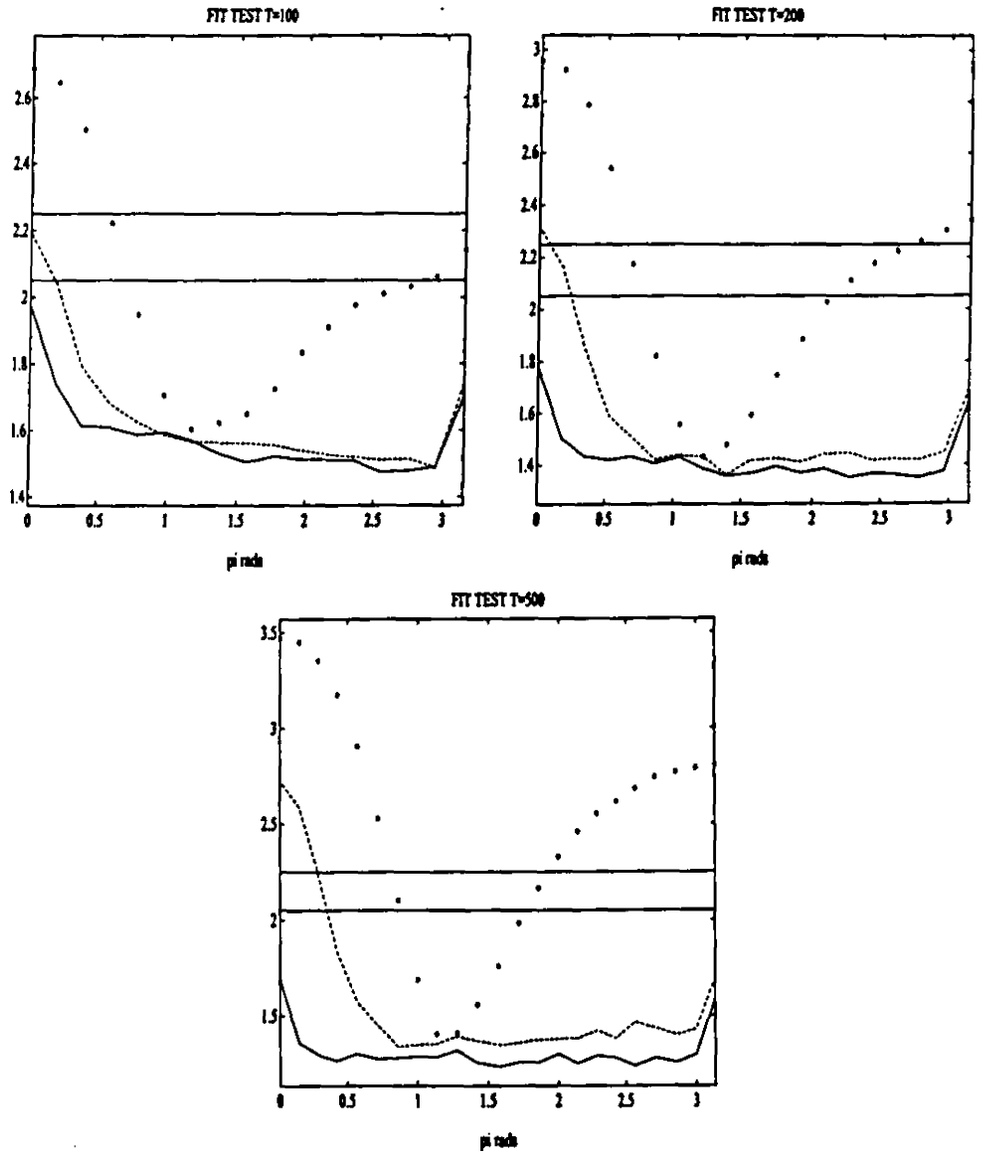

Figure 3.1: **Fit tests for different sample sizes.** Model 1 –, Model 2 - -, Model 3 *. The horizontal lines are the 90% and 95% critical values for the one-frequency test.

Table 3.4: **Parameter values for the IRBC models**

Parameter	Autarky	Trade only	Common shocks	Full Interdep.
Share of Labor in Output (α)	0.64	0.64	0.64	0.64
Growth rate (θ_x)	1.004	1.004	1.004	1.004
Depreciation Rate of Capital (δ_K)	0.025	0.025	0.025	0.25
Discount Factor (β)	0.9875	0.9875	0.9875	0.9875
Steady State hours ($\bar{H}$)	0.20	0.20	0.20	0.20
Risk Aversion (σ)	2	2	2	2
Share of Government Spending in Output (sg)	0.25	0.25	0.25	0.25
Tax Rate (τ)	0.25	0.25	0.25	0.25
Persistence of Technology Disturbances (ρ_A)	0.9	0.9	0.9	0.9
Spillover across Technology Disturbances (ν_{ij})	0	0	0	0.088
Persistence of Government Spending Disturbances (ρ_G)	0.97	0.97	0.97	0.97
Standard Deviation of Technology Innovations (σ_A)	0.00852	0.00852	0.00852	0.00852
Contemporaneous correlation of Technology Innovations (ψ)	0	0	0.258	0.258
Standard Deviation of Government Spending Innovations (σ_G)	0.0036	0.0036	0.0036	0.0036
Imports share (MS)	0	0.15	0	0.15
Armington parameter (ρ)	1.5	1.5	1.5	1.5
Size of each country (Π)	0.5	0.5	0.5	0.5

Table 3.5: **Summary matrix of the fit at business cycle frequencies**

	Autarky	Trade only	Common shocks	Full Interdep.
Autarky	952.9			
Trade only	1612.8	1034.9		
Common shocks	1852.4	2700.6	772.4	
Full Interdep.	1015.2	1457.7	1187.9	932.6
CV 90%	37.9	37.9	37.9	37.9
CV 95%	41.3	41.3	41.3	41.3

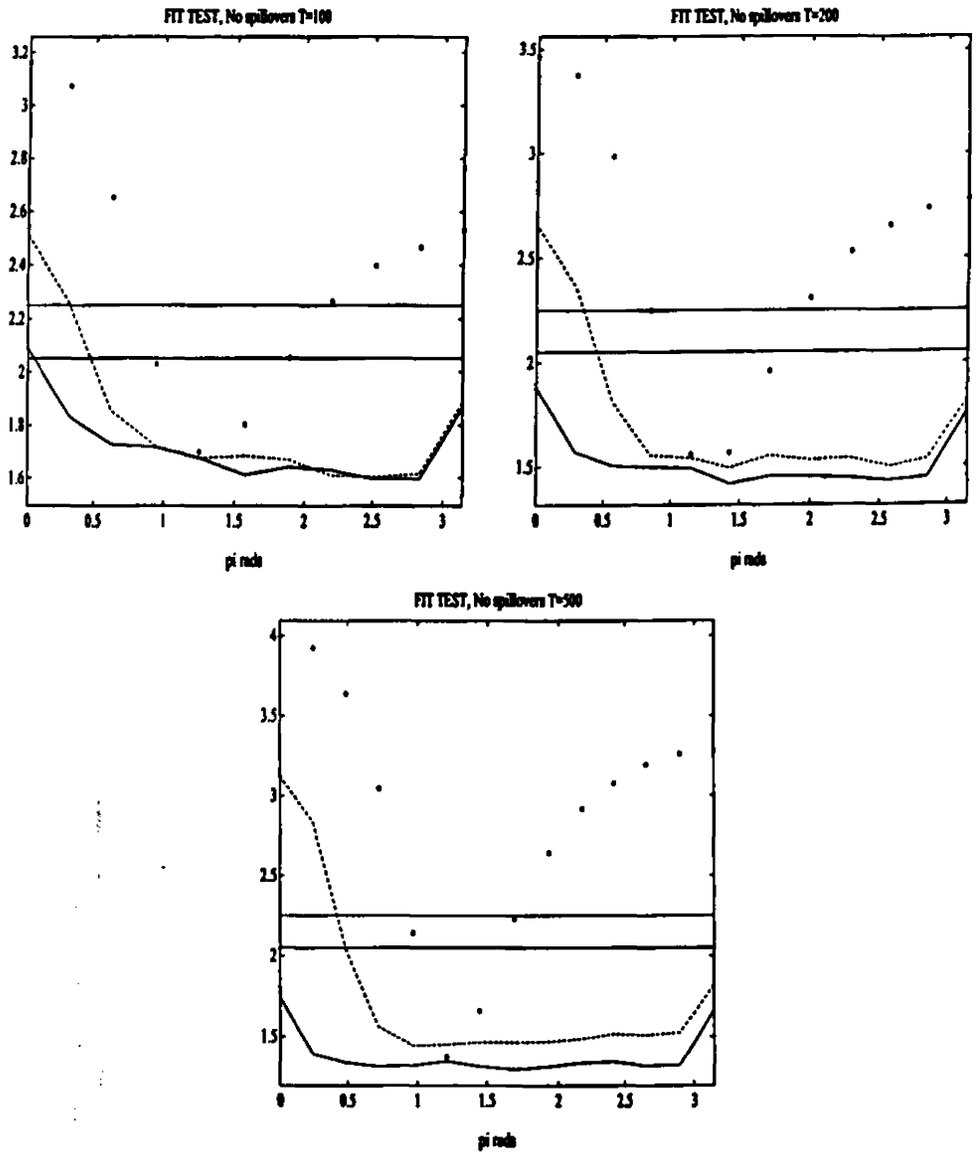

Figure 3.2: **Sensitivity analysis on the Fit test.** Actual data DGP with no spillover among variables (model 2). Model 2 –, Model 1 - -, Model 3 *. The horizontal lines are the 90% and 95% critical values for the one-frequency test.

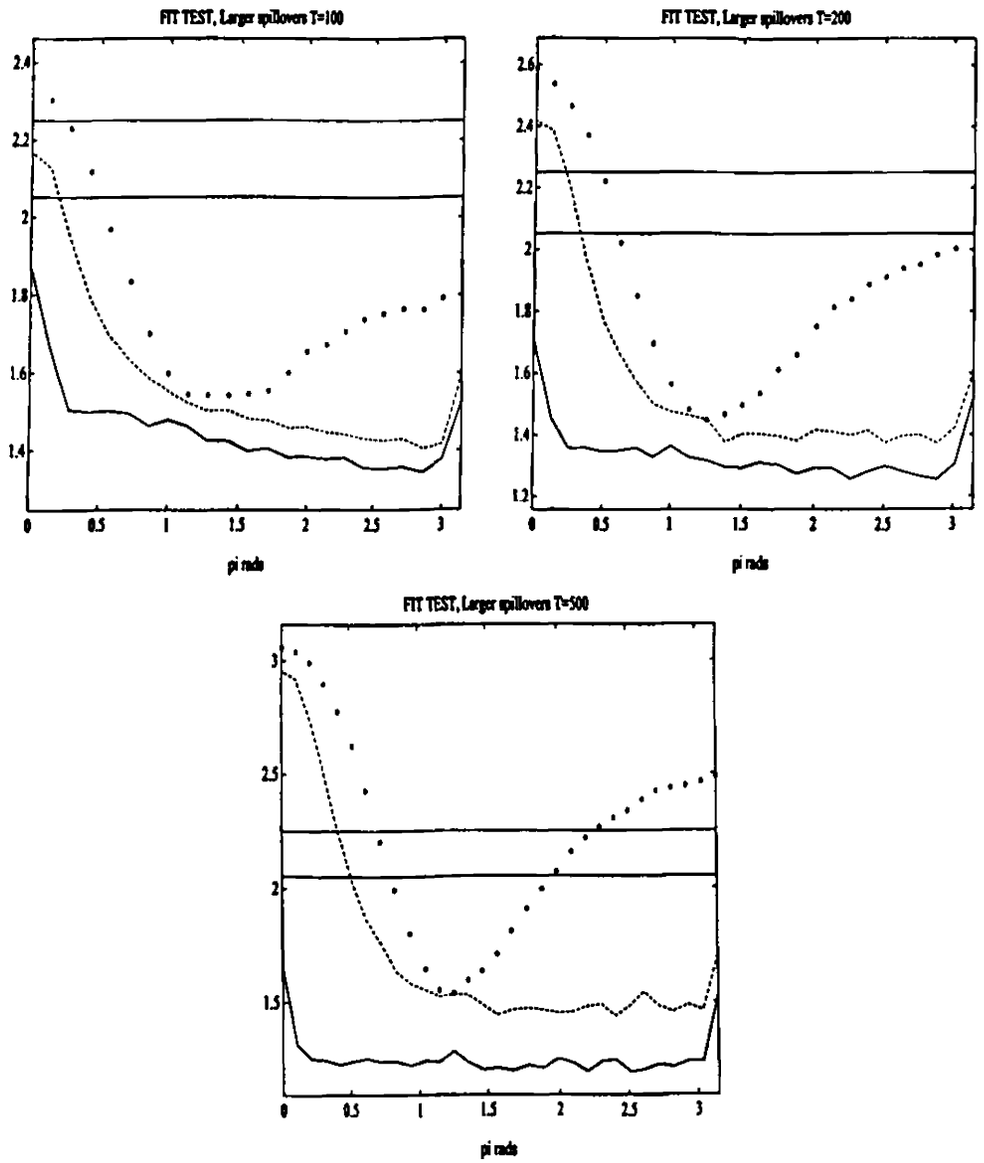

Figure 3.3: **Sensitivity analysis on the Fit test.** Actual data DGP with more pillover among variables. Model equal to actual data DGP –, Model 2 - -, Model 3 *. The horizontal lines are the 90% and 95% critical values for the one-frequency test.

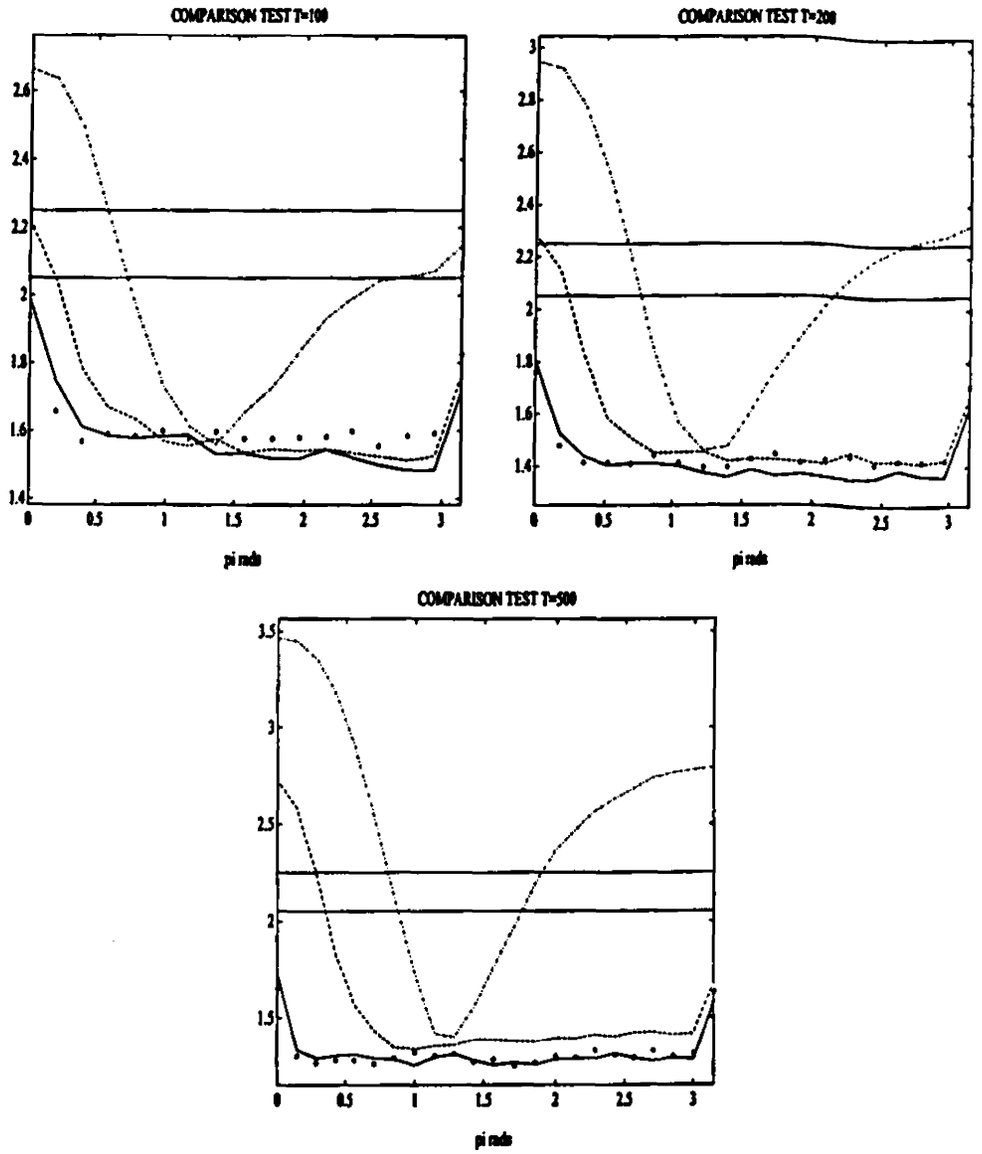

Figure 3.4: Comparison tests for different sample sizes. Case 11 –, case33 *, case 12 - -, case13 -.-. The horizontal lines are the 90% and 95% critical values for the one-frequency test.

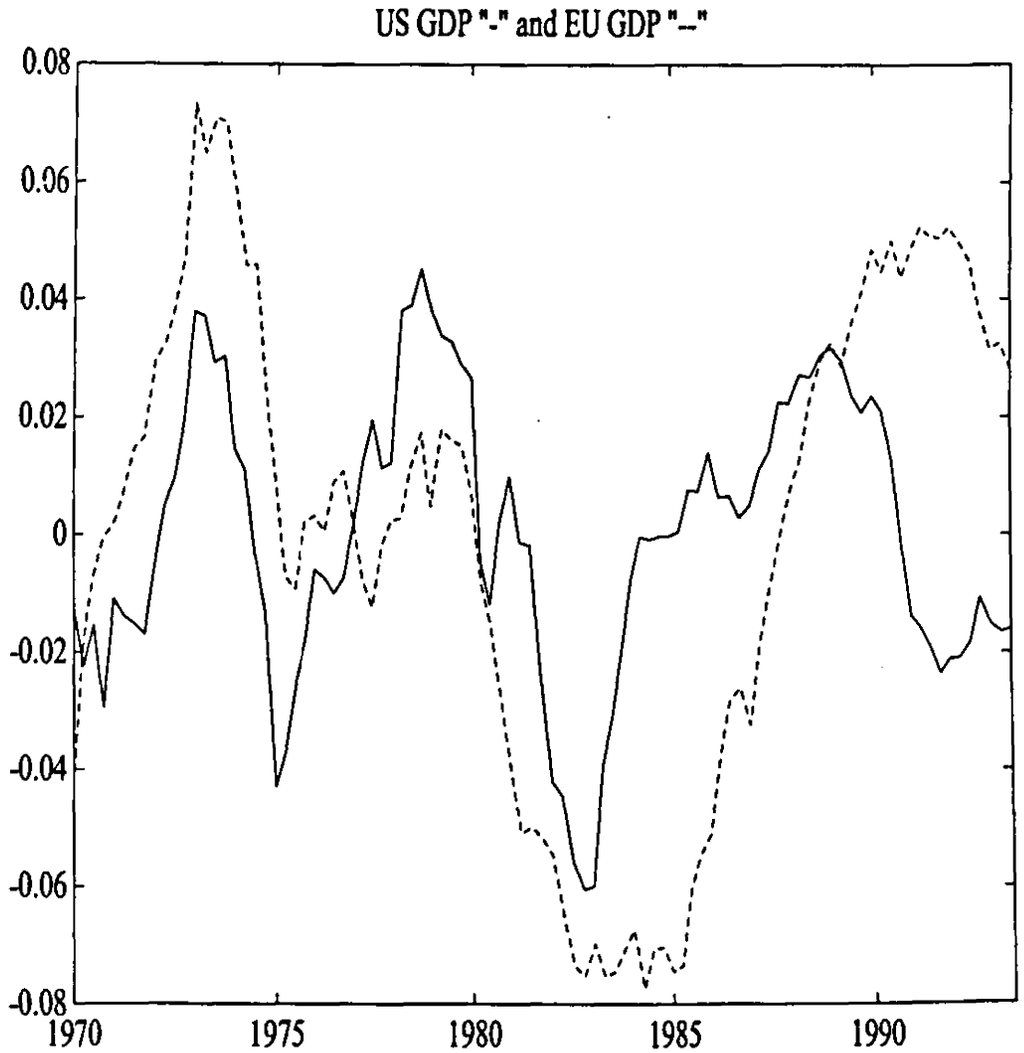

Figure 3.5: **US and European real GDPs, 1970Q1-1993Q3.** Linearly detrended logs of the series.

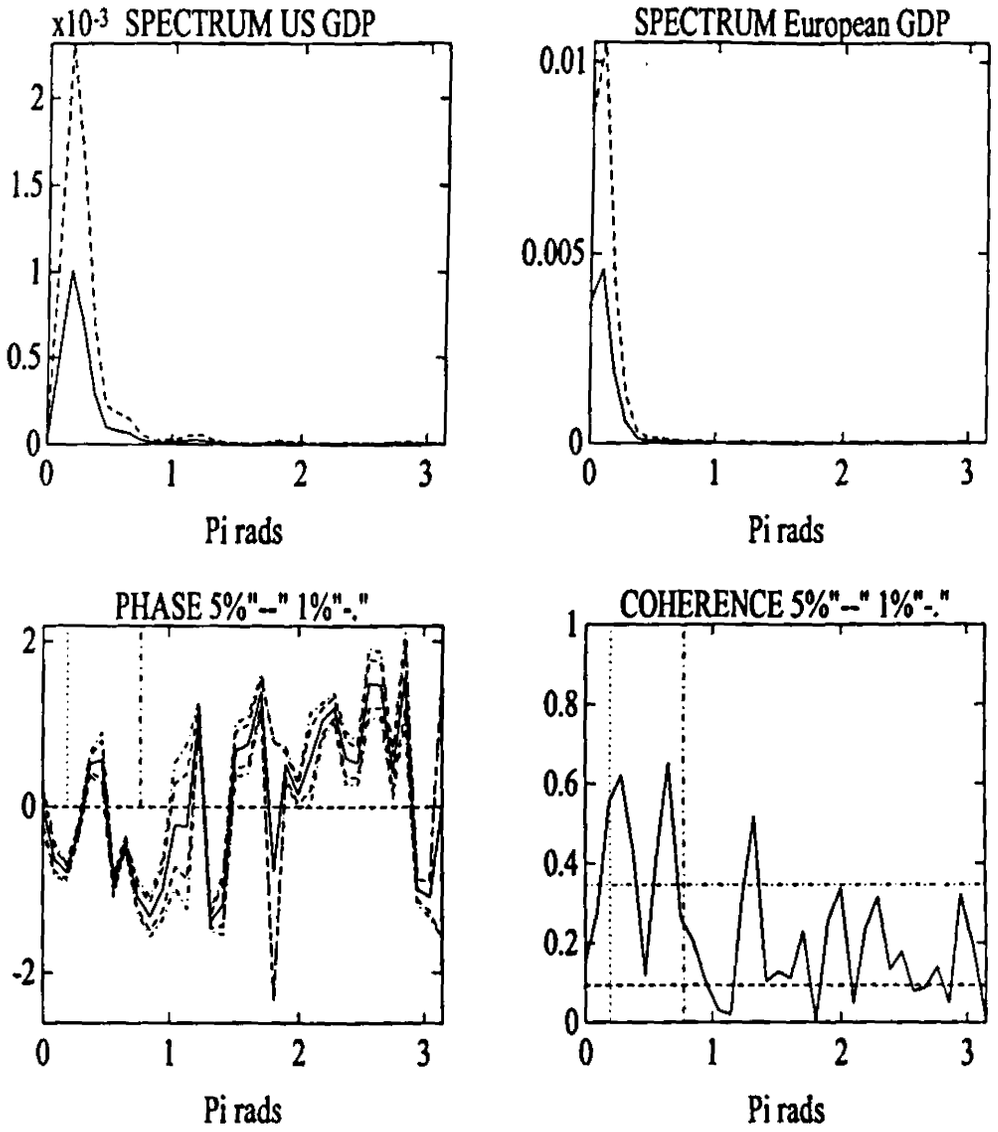

Figure 3.6: **Spectral properties of US and European real GDPs.** Individual spectra in the upper plots (with their 95% asymptotic confidence intervals). The lower plots display the phase (with its 95% and 99% asymptotic confidence intervals –5% and 1% significance levels–) and coherence (with the asymptotic critical values corresponding to the 5% and 1% significance levels, too), with the business cycle interval limited by the vertical lines.

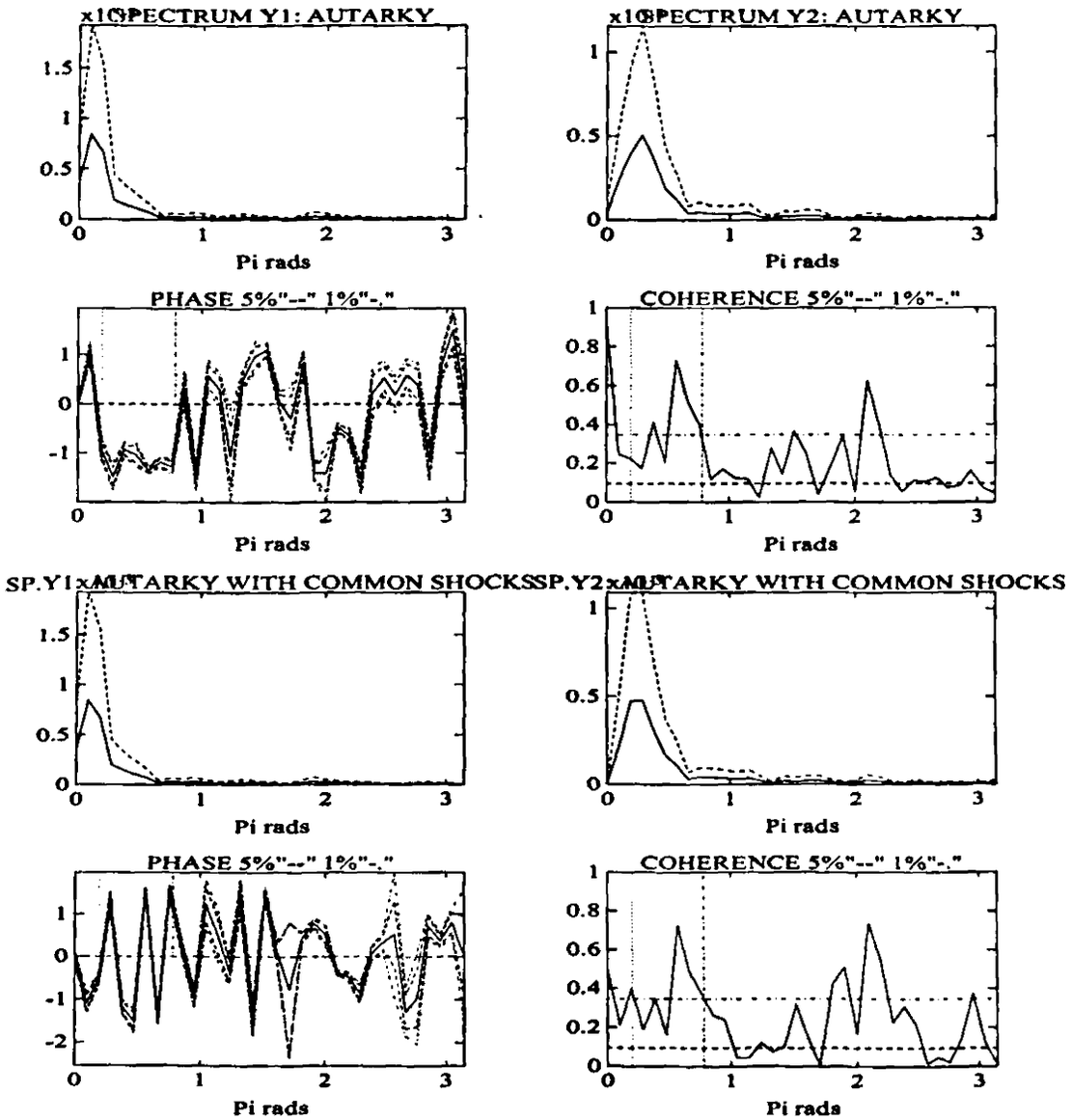

Figure 3.7: **Spectral properties of simulated output series for the two countries** under "Autarky" and "Autarky with common shocks" specifications of the two-country two-good International Real Business Cycle model.

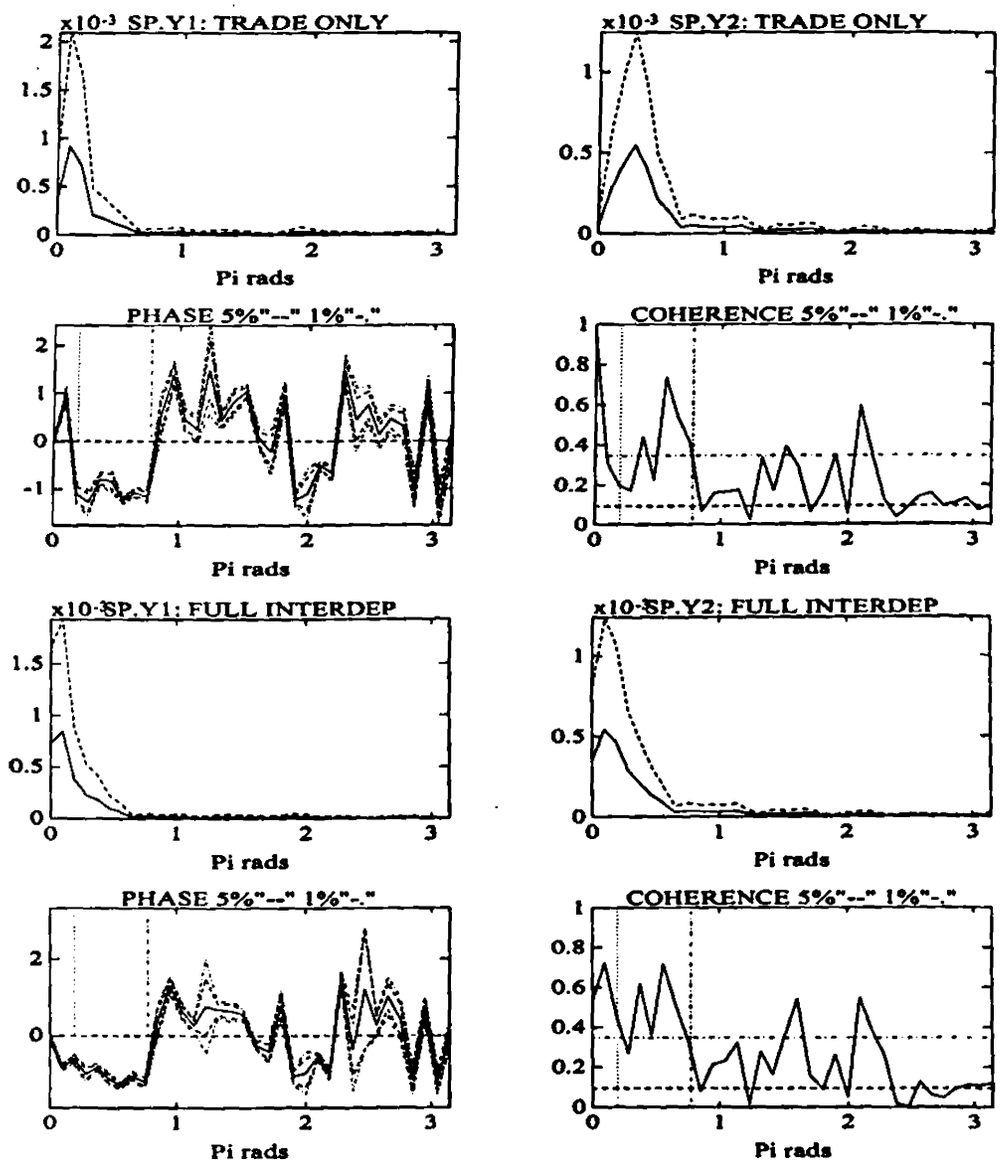

Figure 3.8: **Spectral properties of simulated output series for the two countries** under "Trade Only" and "Full Interdependence" specifications of the two-country two-good International Real Business Cycle model.

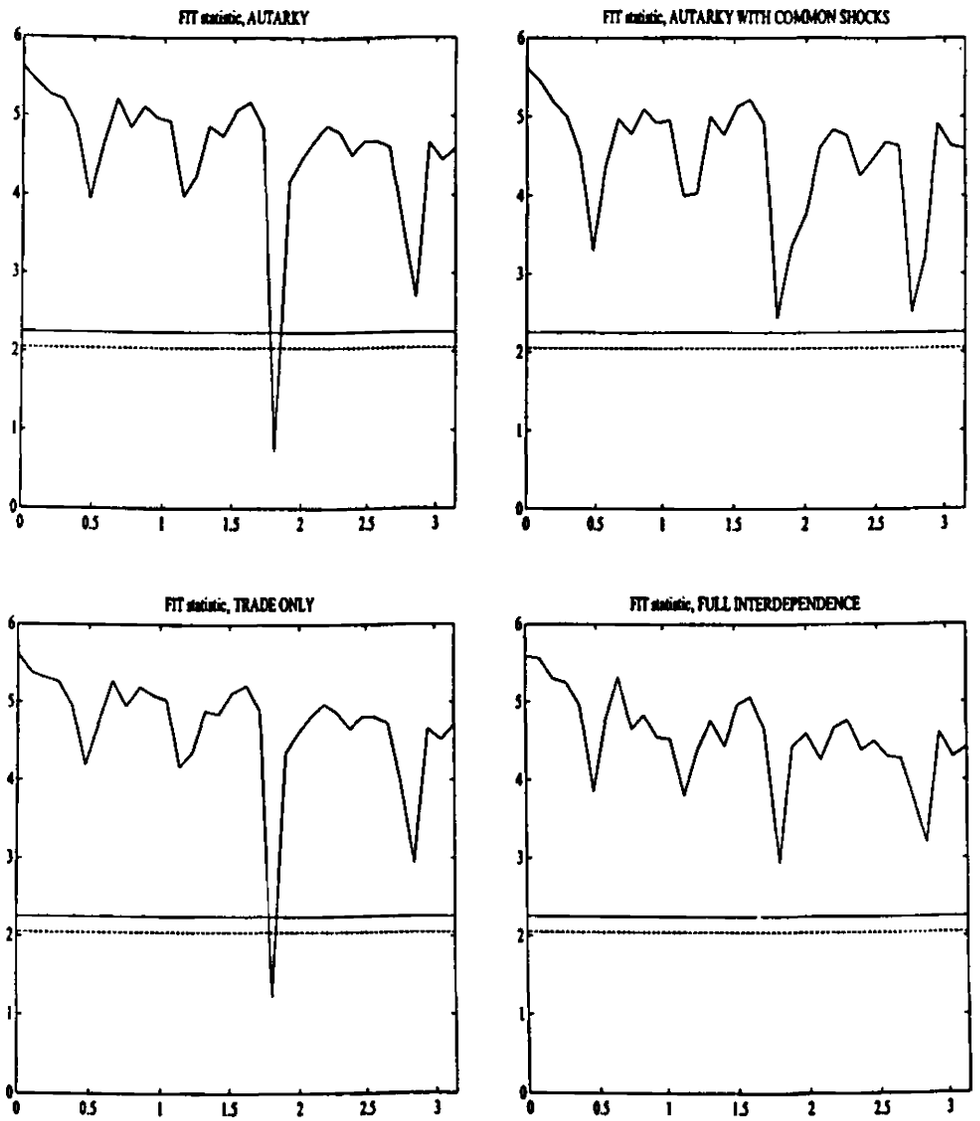

Figure 3.9: **Fit of the four IRBC models.** The horizontal lines are the 90% and 95% critical values for the one-frequency test.

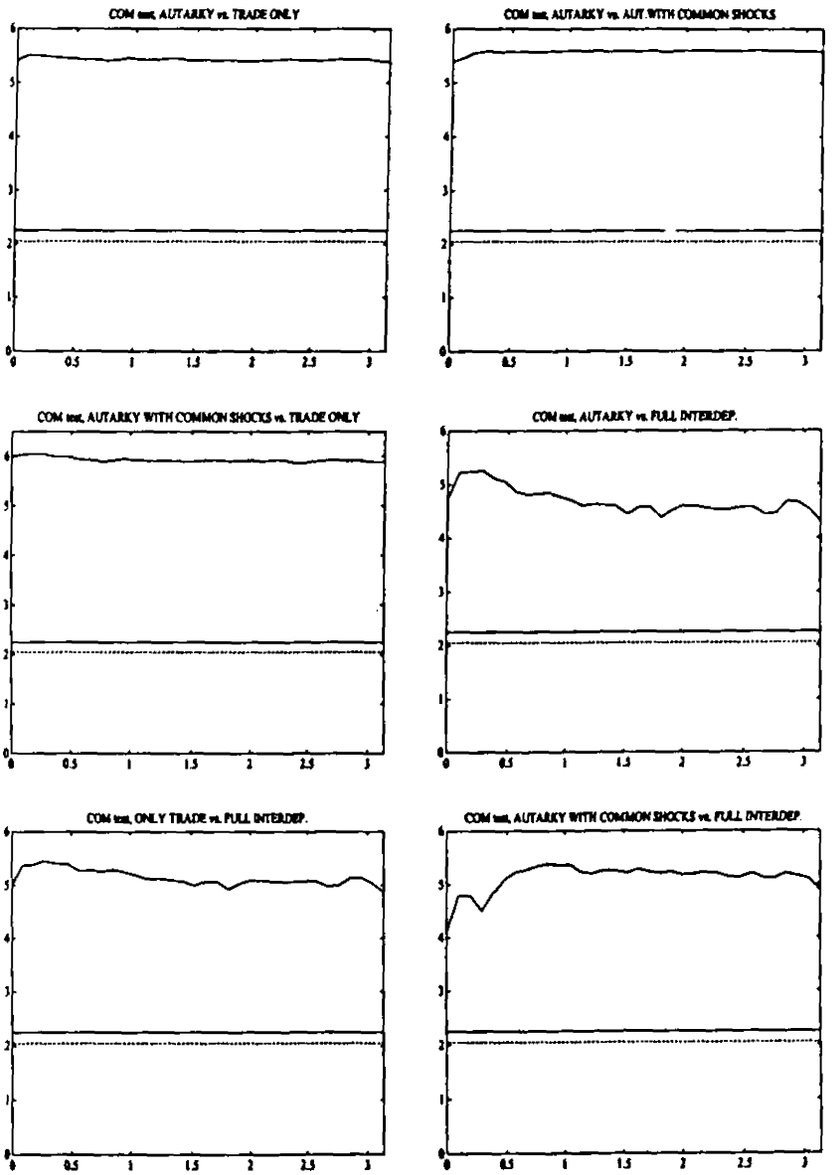

Figure 3.10: **Comparison of the four IRBC models two by two.** The horizontal lines are the 90% and 95% critical values for the one-frequency test.

Chapter 4

Comparing Evaluation Methodologies for Dynamic Stochastic General Equilibrium Models

4.1 Introduction

In the previous chapters we have seen different aspects of the evaluation of calibrated and simulated dynamic stochastic general equilibrium models.

We have overviewed different evaluation methodologies that have been recently proposed to assess the fit of dynamic stochastic general equilibrium macroeconomic models (such as Gregory and Smith (1991), Watson (1993), DeJong, Ingram and Whiteman (1996), Canova (1994), Canova and De Nicoló (1995), Diebold, Ohanian and Berkowitz (1995), among many others), with an emphasis to how simulation-based techniques can be used to formally evaluate the fit of a calibrated model to the actual data. Then we have proposed an alternative methodology for assessing multivariate dynamic models which deals explicitly with the fact that the exact solution of such models is generally not available but approximated.

Each of these model evaluation procedures has been proposed as an alternative to the common informal assessment of a dynamic general equilibrium model which consists on a simple comparison of selected statistics from the model to those of the

actual data. However, it is a difficult task for the researcher to choose which of these model evaluation methodologies to use. Each methodology summarizes the information given by the model in a different way, with different set of statistics, and some of them base their assessment of the model on the distribution of different elements: DeJong, Ingram and Whiteman (1996) use the posterior distribution of the parameters as well as that of the actual data statistics; Canova (1994) uses that of the shocks impinging on the economy and of model parameters; Canova and de Nicoló (1995) use also the distribution of actual data statistics; the fit test suggested in the previous chapter uses the distribution of the spectral density matrix estimator, etc.

This diversity makes it possible that alternative methodologies assess as very different the success of a model in reproducing the same stylized facts of the actual data. A comparison under uniform conditions of the performance of these new methodologies is called for.

This chapter "tests" the performance of a selection of these methodologies using Monte Carlo techniques. It evaluates the ability of each methodology to accept a model when it is equal to the actual DGP and to reject it when it is at odds with the actual DGP. In a sense, we are treating each methodology as a test for dynamic stochastic general equilibrium macroeconomic models and compute its "size" and "power", respectively.

In the next section we describe the experimental design under which we will assess each methodology. We present the benchmark model we will take as the actual DGP in the Monte Carlo experiments and two alternative models. All of them are versions of the King, Plosser and Rebelo (1988) one-sector real business cycle model. Section 4.3 assesses both the benchmark and the two alternative models with respect to actual data for the US informally as it has been done in standard calibration exercises, i.e. based on a simple look at summary statistics from both actual and model data. We try to capture the overall measure of fit of a model under this informal approach with a simple rule which we will refer to as a "naive calibrator" approach.

How do the different evaluation methodologies suggest to improve the informal evaluation of a model by a naive calibrator? And, more importantly, do they effectively lead to more accurate model evaluation than the naive calibrator approach? Are they

only valid under limited assumptions, for evaluating a particular set of statistics or a particular model?

Sections 4.4 to 4.7 answer these questions for four different methodologies: Watson (1993), DeJong, Ingram and Whitemann (1996), Canova and De Nicoló (1995) and the one presented in the previous chapter. For each of them we answer the first question explaining briefly what do they exactly consist on and illustrating them with the assessment of the benchmark and alternative models with respect to actual US data for 1964Q1–1995Q3. The fit of the three models is similar under all methodologies. We also find that results are sensitive to whether model parameters are allowed to vary or not.

The main contribution of this chapter is the answer to the last two questions. For each of the four methodologies studied, as well as for the naive calibrator's rule, we perform a Monte Carlo experiment which "tests" their performance as model evaluation methodologies under uniform conditions.

We find that the naive calibrator's rule we define risks of not accepting even the true DGP and, at best, is only able to reject those models that remarkably differ from the true DGP if the subjective degree of divergence between model and actual data the naive calibrator is willing to tolerate is not appropriately chosen. Watson's approach is found a reasonably accurate methodology. Using a rough measure (only exact for the spectral density distance approach presented in Chapter 2) of the "size" and "power" of each methodology, DeJong, Ingram and Whiteman (1996) and Canova and De Nicoló (1995) appear more accurate than Watson (1993). Among these two approaches, the latter achieves a better "size" at the cost of a lower "power", which is still enough to correctly rank the models according to their discrepancy with the true DGP. The spectral density distance approach is the one which obtains the smaller size and the larger power against models very different to the DGP, but shows no power against real business cycle models similar to the DGP because they generate very similar spectral density matrices to the DGP ones at business cycle frequencies. We find that all four methodologies outperform the naive calibrator's rule since they substantially reduce the risk of rejecting the true DGP, are able to discriminate more clearly between the DGP and models different to it. Section 4.8 concludes.

4.2 The experiment

To compare different model evaluation methodologies under uniform conditions we design the following Monte Carlo experiment: at each replication, we simulate realizations for the shocks impinging on the dynamic stochastic general equilibrium model which we want to evaluate and we draw parameter vectors from their corresponding distribution if they are not fixed but stochastic, compute the statistics of interest from the simulated series each time and assess the model by evaluating how close these statistics are to those generated with the actual DGP (which is known in the Monte Carlo experiment). Model statistics are simulated after generating DGP statistics so that the random numbers do not overlap.

When model parameters are stochastic, the model is simulated several times to introduce in the statistics the uncertainty the researcher associates to the parameters, as in DeJong, Ingram and Whiteman (1996), Canova (1994)-(1995) and Canova and de Nicoló (1995)), but at the cost of a Monte Carlo error since parameters are drawn randomly from their distribution. When model parameters are fixed but the statistics are estimated for model series derived from a realization of the exogenous stochastic disturbances (simulated from their corresponding distribution), we are introducing both a Monte Carlo error (in the simulation of the shock) and an estimation error with respect to using the theoretical statistics implied by the model. Standard calibration exercises typically follow this approach and compute average model statistics across many simulations. The spectral density distance approach presented in Chapter 2 also simulates times series from the model and estimates their statistics. In all cases, errors are reduced if the fit is assessed averaging somehow across many simulations of the model. Typically, around 100 simulations are performed. Watson (1993) uses the theoretical values of model statistics with fixed parameters, in which case no repeated simulation of the model is required.

When illustrating how each methodology works, we will simulate 1000 times the model (except for Watson's approach). But when coming to the Monte Carlo exprimments, the cost of using such a large number of simulations per Monte Carlo replication is too high[7], so we opt for running 100 simulations per replication. We choose to per-

[7]It takes a Pentium around 8 hours to assess the fit of each of the three models we evaluate using

form 100 Monte Carlo replications. Considering that at each Monte Carlo replication we are computing 100 times the simulated statistics, 100 replications is not such a small number, and increasing it would be unfeasible in terms of computer time for some of the methodologies compared.

When the simulated model and the actual DGP are the same any model evaluation methodology should accept the null hypothesis H_0: model = DGP, against H_1: model $\neq$ DGP, except for a predetermined arbitrary α% of the times. The rejection frequency of H_0 across Monte Carlo replications is the empirical "size" of the methodology. When the simulated model differs from the actual DGP then the methodology should reject such H_0 in favor of H_1 in (100-α)% of the replications. The actual percentage rejection of H_0 now is the empirical "power" of the methodology. We would also like the methodologies to reject *more* those models which generate statistics which lay further away from those generated by the actual DGP, i.e. to provide an indication of how far a model is from the actual DGP with respect to alternative models. This point is particularly important for a researcher interested in discriminating between models when none of which is likely to be exactly the actual DGP, which is generally the case for dynamic stochastic general equilibrium models.

The first problem we have to solve is the selection of the actual DGP for the Monte Carlo experiments. We want a model which, on the one hand, highlights the usefulness of all the methodologies studied and, on the other hand, allows to derive easily alternative models that differ from it in different degrees so that we can measure how each methodology captures the degree of divergence between each alternative model and the actual DGP. Some of the methodologies we compare (Watson (1993), DeJong, Ingram and Whiteman (1996) and Canova (1994)-(1995)) are specific for assessing calibrated models, Canova (1994)-(1995), Canova and De Nicoló (1995), DeJong, Ingram and Whiteman (1996) and the fit test proposed in the previous chapter apply to simulated dynamic stochastic general equilibrium models and the fit test is constructed for assessing multivariate dynamic models whose solution has to be approximated. Hence, we select a calibrated dynamic stochastic general equilibrium model whose solution is not exact but approximated and which can generate simulated statistics of interest

the DeJong, Ingram and Whiteman (1996) methodology, not much less using Canova and De Nicoló (1995) methodology, and around 12 hours using the spectral density distance approach.

sufficiently different from each other when generated under different hypothesis.

It is hard to compare different methodologies when there is not much room for the alternative models to generate different statistics since no evaluation methodology will be able to recognize how relatively different alternative models are. Instead, a simple model whose generated time series are very sensitive to the particular model specification would be the best choice. A very well known prototype of such model is the King, Plosser and Rebelo (1988) one-sector economy with government spending and technology shocks. Such a model has the further advantage that some of the methodologies we compare in this chapter have been applied to versions of this model so we can have in some cases direct means of comparison in the literature.

Next we have to select with respect to which particular features of the data we want to assess the fit of the model. We choose to focus on few multivariate statistics: the relative standard deviation between per capita consumption (C) and output (Y) and the contemporaneous correlations between C and Y, hours (H) and output, and hours and average labor productivity (AP). When the model evaluation methodology is performed in the frequency domain the standard deviation is replaced by the power spectrum and the correlation by the coherence at selected frequencies. The reason for selecting these multivariate relationships is that they include some statistics which are typically successfully captured by the different versions of the model (e.g. corr(Y,C)), some others which typically are not (e.g. corr(H,AP) is typically too high and std(C)/std(Y) too low in the model with respect to the actual data) and some which vary across model specifications (e.g. corr(Y,H) is generally too high in some models but very close to the actual data in others). Hence, there is room to generate different statistics under alternative specifications of the model.

The rest of this section presents briefly the model used as the actual DGP in the Monte Carlo experiments (Model 1) and its statistical implications and two versions of this model (Model 2 and Model 3) which will be used to check the power of the various evaluation procedures.

4.2.1 The models

The three models we consider are versions of the basic real business cycle model explained in detail in King, Plosser and Rebelo (1988): the one-sector neoclassical model of capital accumulation where work effort is a choice variable and economic fluctuations are initiated by impulses to technology or by shocks to government spending.

Model 1 has technology shocks as the only source of economic fluctuations, is the most commonly studied one-country real business cycle model and has been used in the literature for illustrating some of the new model evaluation methodologies we are comparing in this chapter (Watson (1993) and DeJong, Ingram and Whiteman (1996)). Model 2 includes government spending shocks. The addition of this further shock alters the dynamics while keeping similarities with the benchmark model. Model 3 allows for government spending shocks only and generates very different model dynamics. In what follows, we present first the model structure and its solution and afterwards we specify the parameterization for each model and their implications for our multivariate statistics.

The economy is populated by a large number of identical infinitely-lived agents. All variables are expressed in per capita terms. Preferences of the representative agent are given by:

$$U \equiv E_0 \sum_{t=0}^{\infty} \frac{\beta^t}{1-\sigma} C_t^{1-\sigma} v(L_t) \tag{4.1}$$

where C_t is private consumption of the single good by the representative agent and L_t is leisure, β is the discount factor and σ the coefficient of relative risk aversion. Leisure choices are constrained by:

$$0 \leq L_t + H_t \leq 1 \tag{4.2}$$

where the total endowment of time is normalized to 1 and H_t represents the proportion of time devoted to market activities.

The single final good is produced with a Cobb-Douglas technology:

$$Y_t = A_t (K_t)^{1-\alpha} (X_t H_t)^{\alpha} \tag{4.3}$$

where K_t is the capital input, α is the share of labor in GDP, and X_t is labor-augmenting Harrod-neutral technological progress with deterministic growth rate equal to θ_x, i.e.

$X_t = \theta_x X_{t-1}$ with $\theta_x \geq 1$. X_t represents permanent technical change while temporary changes in technology are represented by variation in total factor productivity according to

$$lnA_t = \rho_A lnA_{t-1} + \epsilon_{At}$$

where $\epsilon_{At} \sim N(0, \sigma_A^2)$.

Capital goods are accumulated according to:

$$K_{t+1} = (1 - \delta_K)K_t + I_t \tag{4.4}$$

There is an output tax whose revenues are used to finance an exogenous path of per capita government expenditures G_t and lump sum transfers TR_t. These expenditures are assumed not to affect the economy's production possibilities nor the representative agent's marginal utility. The government budget constraint is

$$G_t = TR_t + \tau Y_t \tag{4.5}$$

where G_t follows the stochastic process:

$$G_t = \rho_G G_{t-1} + \epsilon_{Gt}$$

where $\epsilon_{Gt} \sim N(0, \sigma_G^2)$. Innovations to total factor productivity, ϵ_{At}, and to government spending, ϵ_{Gt}, are assumed to be independently distributed.

The economy wide resource constraint is given by:

$$Y_t - G_t - C_t - I_t \geq 0 \tag{4.6}$$

All variables except hours and leisure are assumed to grow in the steady state at the same rate as the technological progress, θ_x-1, so that business cycle dynamics are separated from growth associating the latter to that deterministic trend common to all drifting variables. Technology, preferences and government behavior are restricted following King, Plosser and Rebelo (1990) so that the suboptimal (because of distorting taxes) competitive equilibrium solution is compatible with steady state growth. The equilibrium solution is characterized by the efficiency conditions for the individual's maximization problem together with the government constraint.

To characterize the local dynamics around the steady state path, i.e. what happens to the economy when it faces alternative sequences of exogenous shocks, we follow King, Plosser and Rebelo (1990) and express the transformed efficiency conditions in terms of detrended variables: we take ratios of the original per capita drifting variables with respect to the labor augmenting technological progress so that the economy is transformed from steady state growth to stationarity. The modified optimality conditions are then approximated with a log-linear expansion around the steady state.

Time series for consumption (C), output (Y), hours (H) and average labor productivity (AP) are generated from the approximate optimality conditions, once the free parameters and time series for the innovations to exogenous processes of the model are given. The statistics of interest (standard deviations and correlations or spectra and coherences) of simulated data are computed after extracting from the raw simulated time series a linear trend, in the same fashion as actual data statistics.

Table 4.1 shows the parameter values for each of the three model specifications we consider. They only differ in the parameterization of the exogenous processes so that Model 1 (the actual DGP in the Monte Carlo experiments) has only technology shocks, Model 2 has both technology and government spending shocks and Model 3 has only shocks to the exogenous government spending process.

Parameter values are taken not only from King, Plosser and Rebelo (1988) but also from literature related. We have estimated θ_x as King, Plosser and Rebelo (1988) do, one plus the average quarterly rate of growth of real per capita output, but with an updated data set (1964Q1–1995Q3 instead of 1948Q1–1986Q4). σ=2 is the standard value calibrated models use for multiplicatively separable momentary utility. We impose government budget balance in the steady state by assuming a constant tax rate (τ) equal to a constant government spending output share (sg). We have taken a value for τ and sg which lays in between the one suggested by King, Plosser and Rebelo (1988) of 30% and that used by Baxter and King (1993) of 20% for the case of steady state balanced budget (Aiyagari, Christiano and Eichenbaum (1992) suggest a government spending share of 17.7%). σ_A is the standard value used in the literature for the standard deviation of technology innovations. The persistence of government spending process (ρ_G) and the standard deviation of its innovations (σ_G) are from Aiyagari,

Christiano and Eichenbaum (1992).

The first three columns of Table 4.2 show the statistics of interest for each of the three models (see Stadler (1994) for a good summary of the basic implications of real business cycle models).

Positive temporary technology shocks increase output and hence consumption, generating positive and large consumption-output contemporaneous correlation in Models 1 and 2. Consumption increases to a lesser extent since agents seek to smooth it over time (so that the relative standard deviation of consumption with respect to output is lesser than one in all three models) and this increases the capital stock. Temporary productivity shocks shift the production function and hence the labor demand curve. The marginal product of labor is also increased, but since the utility function is specified so that the income and substitution effects of a real wage change cancel each other, the labor supply curve does not shift. Then, under Model 1 and 2 the shift in labor demand increases real wages[8] and the hours worked. Because of the intertemporal substitution of leisure these models generate high positive hours-output contemporaneous correlation and significantly positive hours-average product of labor one (this last observation is referred to in the literature as the "productivity puzzle").

Government spending does not enter directly the agent's utility function (nor the production function) and hence shocks to government expenditure do not have a substitution effect but only a wealth effect. That is the reason why when we added them to technology shocks as the sources of business cycle fluctuations (Model 2) the statistics displayed in Table 4.2 do not change that much (correlations are slightly reduced). Things change, though, when government spending shocks are the only source of fluctuations (Model 3). The rise in government spending financed by taxes results in a negative wealth effect that shifts the labor supply curve while the labor demand one remains unchanged. This produces an increase in hours worked (contemporaneous correlation with output of 1) and a decrease in real wages (contemporaneous hours-average labor productivity of -1). Government consumption crowds out consumption through this negative wealth effect, resulting a contemporaneous consumption-output

[8]The Cobb-Douglas specification of the production function implies that the marginal product of labor (real wage in competitive equilibrium) will move quite closely with the average product of labor, or productivity. Therefore, the increase in real wages is translated into an increase in AP.

correlation of -1.

4.3 An informal evaluation

Table 4.2 shows also the statistics of the actual data. Data is from OECD Quarterly National Accounts, in constant 1985 US$Bln, and from National Government OECD Series (Department of Labor) in thousands of people, all seasonally adjusted. The sample period is 1964Q1–1995Q3. Y is GNP, C is personal final consumption expenditure, H is total civilian employment times average weekly hours of all private workers on nonfarm payrolls. Variables are transformed into per capita terms dividing them by civilian noninstitutional population of 16 and more years old excluding armed forces (source: Department of Labor, National Government OECD Series). AP is Y/H. To maintain a close relationship between the model and the actual data we linearly detrend the logs of all the series but for H (and AP is detrended by substracting log(H) from the detrended log(Y)) before computing the four statistics we are interested in. The statistics differ slightly from those reported in other works for two reasons: because we have used an updated data set and also because the detrending method chosen has an impact on these statistics (see Canova (1997)).

Informal evaluation of how the three models reproduce the relationships between output, consumption, hours and average productivity observed for US data would consist on casual inspection of columns 1, 2 or 3 of Table 4.2 compared to the last column. The conclusion would be that Model 1 and 2 are similarly successful in reproducing the observed facts although they predict too little consumption variability and too much hours-productivity contemporaneous correlation. Instead, Model 3 would be rejected as a good explanation of the observed facts since it totally misses the positive high consumption-output correlation and predicts an hours-productivity correlation of -1.

Very often dynamic stochastic general equilibrium models are judged successful or rejected according to similar informal criteria. An informal evaluator would consider the model more successful than others the larger the number of model statistics which are similar to the actual data ones and the smaller the divergence between actual and simulated statistics. To put this formally we arbitrarily choose the following rule: a

"naive calibrator" would reject a model if at least 3 out of the 4 statistics he is interested in explaining differ in absolute value from the observed ones by more than x%.

We perform the Monte Carlo experiment outlined in Section 4.2 on this naive calibrator's rule to evaluate how reliable such an informal criterium is to accept or reject a model. We compute the rejection frequencies of the null hypothesis H_0: Model i = DGP, against H_1: Model $i \neq$ DGP, for i=1, 2 and 3. At each Monte Carlo replication, we simulate 100 times time series of the usual length (127 observations as we had for actual US data) from a model and use the naive calibrator's rule to compare the average (across simulations) of the 4 statistics we are interested in to the actual DGP statistics. Model 1 is taken as the DGP so that DGP statistics are the theoretical statistics of Model 1 (calculated using simulated series of 10,000 observations), which are kept fixed across Monte Carlo replications and across experiments. If the null hypothesis H_0: Model 1 = DGP is rejected 0% of the times and the rejection frequencies of H_0: Model 2 = DGP and of H_0: Model 3 = DGP are 100%, the naive calibrator's rule is a perfect model evaluation methodology (0% *size* and 100% *power*) since it is always able to recognise which are the correct and incorrect models. Obviously, these rejection frequencies will depend on the x% the naive calibrator is willing to consider as a significant divergence between actual and model statistics.

Over 100 replications, we find that Model 1 is rejected 0% of the times when the rule is: reject if the divergence exceeds 50% of the actual data statistic for 3 or more out of 4 statistics. Such a "permissive" rule leads to reject H_0: Model 2 = DGP also 0% of the times while rejects H_0: Model 3 = DGP 100% of the times. Being so permissive, the naive calibrator will always consider as equal to the DGP models which are not equal but similar to the DGP (such as Model 2). However, reducing the accepted divergence to 10% of the value of the actual data statistics, the naive calibrator's rule becomes "too strict", in the sense that although it succeeds to reject models different to the DGP (Model 2 and Model 3) 100% of the times, it also rejects the true model (Model 1) 100% of the times. This is because with short time series and using the average statistics (across simulations) induces a large enough error which makes some simulated statistics differ in more than 10% from the true ones. The following table shows the average across Monte Carlo replications of the divergence between model simulated

statistics (averages across 100 simulations of the model) and actual DGP statistics (theoretical statistics of Model 1), measured in percentage of the values of actual DGP statistics. We are actually computing for each statistic the x% divergence which, on average, should be allowed by the naive calibrator in order to accept the H_0: Model i = DGP. The results of the first column are striking, i.e. the divergence between the theoretical statistics implied by Model 1 (actual DGP) and those computed averaging statistics from short series generated also from Model 1 is very large.

Statistic	Model 1:	Model 2:	Model 3:
std(C)/std(Y)	28%	21%	14%
corr(C,Y)	4%	5%	211%
corr(H,Y)	57%	56%	92%
corr(H,AP)	423%	398%	135%

A naive calibrator with a rule not accepting less than 28% divergence between model and actual data statistics for 3 or more out of 4 cases would not be able to accept the true DGP (i.e. the naive calibrator's decision rule would have a huge "size"). Whereas one would need to accept a percentage divergence as high as 92% for 3 out of 4 statistics to accept models having very different equilibrium dynamics than the true DGP such as Model 3 (i.e. the naive calibrator's rule would have no "power").

Summarizing, the naive calibrator's rule we have defined risks of not accepting even the true DGP and, at best, is only able to reject those models that remarkably differ from the true DGP if the subjective degree of divergence between model and actual data the naive calibrator is willing to tolerate is not appropriately chosen. The results of this Monte Carlo experiments strongly advice dynamic stochastic general equilibrium modellers not to rely on averaging, across several simulations of the model, the statistics of short simulated time series, as it is often found in the literature of calibrated models.

How do more formal evaluation methodologies suggest to improve the evaluation of a model by a naive calibrator? And, more importantly, do they effectively lead to more accurate model evaluation than the naive calibrator approach? Are they only valid under limited assumptions, for evaluating a particular set of statistics or a particular model? The following sections answer these questions in three steps. First, we describe

briefly each methodology. Second, we illustrate how they work by evaluating our three models. Finally, we check their performance as model evaluation methodologies with the Monte Carlo experiment described, in the same fashion as we have done to the naive calibrator's decision rule.

4.4 Watson's measures of fit

Watson (1993) suggests a way to evaluate calibrated models by asking how much error should be added to a model generated series, $x_t^* = g(z_t, \gamma)$ (where γ are the parameters of the model and z_t are exogenous stochastic disturbances) so that its spectral density equals the spectral density of the corresponding actual data series y_t. As explained in Chapter 1, the error $u_t^* = y_t - x_t^*$ includes both model error ($u_t = y_t - x_t$) and the error of approximating with x_t^* the exact model solution $x_t = f(z_t, \gamma)$ since it is most of the times not obtainable analytically. The choice of the spectral density function of y_t as the set of stylized facts of the data to be matched by the model has clear advantage over selecting relative standard deviations and correlations when we are interested in evaluating the business cycle properties of a model, because we can focus easily on only those frequencies associated with business cycle fluctuations.

Watson provides an R^2-type measure of fit between the model and the data based on the ratio of the spectral density of the error $A_{u^*}(\omega)$ to that of the actual data $A_y(\omega)$ for a particular frequency ω or for a frequency range $[\omega_1, \omega_2]$ (see Chapter 1). The size of the ratio is evaluated informally (i.e. whether it is greater than one, between zero and one or close to zero). This ratio is a lower bound: when it is large the model poorly fits the data but when it is small it does not necessarily follow that the model is appropriate. Note that in this approach, γ and z_t are fixed, and A_{x^*} and A_y are assumed to be measured without error.

Watson's measures of fit (for a single frequency or for a frequency range) are univariate but can be easily extended to a multivariate evaluation of a model in the same fashion as has been done in the example in Chapter 1, so that we can evaluate how well our three models reproduce the multivariate relationships between Y, C, H and AP observed in US data for 1964Q1-1995Q3. Table 4.3 reports the results of such evaluation. All statistics reported in Table 4.3 are averages across business cycle frequencies, i.e.

those associated to cycles 2 to 8 years long.

We have estimated the spectral density matrix for the linearly detrended 4-variable actual data set as well as that implied by each model[9]. Spectral density matrices are estimated using a Bartlett window (see Priestley (1981), ch.9.5) with a sample size-dependent bandwidth parameter $M = 1 + 3 \times T^{1/3}$ so that we capture the optimal rate of convergence of the Bartlett window of $O(T^{1/3})$ (see Andrews (1991)). Figures 4.1 and 4.2 display spectra and coherences for actual US data and for the model series generated under Model 1, 2 and 3, and for all frequencies.

We compute Watson's measure of fit for each of our four series

$$R_j = \frac{\int_{\omega \in Z} A_{u^*}(\omega)_{jj} d\omega}{\int_{\omega \in Z} A_y(\omega)_{jj} d\omega}, \quad j = 1,2,3,4$$

where $A_y(\omega)_{jj}$ and $A_{u^*}(\omega)_{jj}$ are the actual data and the error spectral densities, respectively, for series j and where the ω frequencies included in the Z interval are business cycle frequencies (see section 1.3 in Chapter 1 for a more detailed explanation). R_j is calculated under two different identification schemes: one which minimizes the trace of A_{u^*} with equal weight to its 4 components, and a second one which minimizes the spectral density of the error associated to output (when there is only one source of fluctuations: technology shocks in Model 1 or government expenditure shocks in Model 3) or to output and hours (when there are two types of shocks in the model, i.e. Model 2). The measures of fit for the consumption-output, hours-output and hours-average product of labor coherences (i.e. the frequency domain equivalents to corr(C,Y), corr(H,Y) and corr(H,AP), respectively) are calculated as the ratio between the average coherence of model series across BC frequencies and that of actual data coherence, since it is hard to interpret what the coherence between approximation errors means. Instead, such measure is expected to be closer to 1 when the observed coherence is explained

[9]Instead of deriving the theoretical spectral density of each model as in Watson (1993), we have estimated it for the 4 variables simulated using the parameter values of Table 4.1 and simulating the model only once. We have simulated time series 1000 observations long so that their estimated model spectra are sufficiently close to their theoretical values. Sensitivity analysis has been performed on the effect of the model series length, i.e. all the statistics in Table 4.3 have been computed for the case in which model spectral density is estimated from series 500, 200 and 100 observations long. The main result is that the measures of fit statistics increase in general the shorter the model series length, indicating a worse fit.

by the model. Note that, by construction, such a statistics it is not affected by the identification scheme chosen.

Table 4.3 shows, consistently with the informal inspection of the relative standard deviation and correlations of Table 4.2, that the fit of Models 1 and 2 is very similar and much more satisfactory than that of Model 3 (only government spending shocks). It changes across identification schemes and seems to be better when equal weights are given to all 4 approximation errors. The advantage of Watson's approach over the informal evaluation is that it allows us to know the percentage of the spectral density of each actual series that the model is missing, which ranges from 2.7% for Y to 31% for AP in Model 1 (2.6% and 28% in Model 2) but from 91% for Y to 105% for C in Model 3. The coherence between C and Y is particularly well captured by Models 1 and 2 (only 5% and 2% higher than in actual US data respectively) and not that bad by Model 3 (27% higher in the model), while the hours-AP coherence is particularly badly captured in all three models. In general, Watson's measures of fit lead to prefer Model 2 to Model 1 (they have lower values) and to reject clearly Model 3 as possible explanations of the business cycle relationships between Y, C, H and AP observed in the US in 1964Q1-1995Q3.

4.4.1 Evaluating Watson's approach

Next we face Watson's methodology with a "test" similar to the one faced by the naive calibrator's rule. At each replication of the Monte Carlo experiment, we keep the theoretical spectral density matrix of the model fixed across Monte Carlo replications and estimate the actual data spectral density matrix. Actual data being generated from the DGP (Model 1), we simulate once series of a length usually found in practice (we actually take 127 observations as we had for the US data) using the parameter values of the first column in Table 4.1 and estimate with a Bartlett window their spectral density matrix. Then, the 7 measures of fit corresponding to BC frequencies are calculated minimizing the spectral density matrix of the approximation errors according to one of the two identification schemes explained above. Table 4.4 summarizes the empirical

distribution of the 7 measures of fit across the 1000 replications[10] for each model being evaluated and for each identification scheme, with the mean, the standard deviation, and the 5%, 50% and 95% percentiles.

The median measure of fit across Monte Carlo replications indicates an error of .07% (cohe(H,Y)) to 7.4% (sp(C)) for Model 1, when 0% should be expected. This can be considered the "size" of Watson's model evaluation methodology according to our Monte Carlo experiment. As we pointed out when evaluating the naive calibrator's rule, apart from the obvious Monte Carlo error, this error may come mainly from the fact that we are comparing model spectra estimated for short time series simulated from the DGP to the theoretical model spectra. However, the divergence using Watson's approach is considerably smaller than that we observed for the naive calibrator's approach[11].

When assessing Model 2, the median measure of fit ranges from 2% (cohe(H,Y)) to 7.3% (sp(C)). This can be considered a measure of the "power" of Watson's approach versus models which are known to be close to the actual DGP. Such values indicate a worse fit (are further away from 0% for the spectra and from 100% for the coherences) for Model 2 than for Model 1, as we would expect. In both cases, the measures of fit increase when the identification scheme weights differently the errors (we obtain a range of median measures of fit of .07% to 38% for Model 1 and of 2% to 69% for

[10]We choose 1000 replications for two reasons: first, because increasing the number of replications would have had a big cost in terms of computing time since we are computing the Monte Carlo distribution of the Watson measure of fit twice for each of our three models and, second, because there cannot be less replications if we want the results to be comparable to other methodologies (for which we perform only 100 replications but model statistics are simulated 100 times per replication).

[11]We have performed a further sensitivity analysis on this issue and computed Table 4.4 for the case in which model spectral density matrices are also estimated for short time series simulated from the corresponding model instead of taking their theoretical values. That is, instead of simulating model series of 1000 observations, we have estimated the spectral density matrix from model series of length 500, 200 and 100.

For model series of 100 observations, the range of the median measure of fit rises to 4% to 21% when testing H_0: Model 1 = DGP (the range is equal under both identification schemes but most values are lower when equal weights are taken) but is 4.3% to 8% (8% to 54% when weighting only Y and H approximation errors) when testing H_0: Model 2 = DGP. The error induced using statistics estimated for short simulated time series leads to prefer Model 2 to the true DGP which is Model 1. The median mesures of fit when testing H_0: Model 3 = DGP for simulated series of 100 observations range from 17% to 104% (111% when weighting only Y errors) which leads to reject Model 3 as clearly as when using the theoretical model spectral density matrix.

Model 2). Although the "size" gets worse, the "power" to discriminate between the actual DGP and other models gets better. The bottom part of Table 4.4 shows a high "power" against models very different to the DGP. The median measures of fit indicate an error ranging from 16% to more than 100% under both identification schemes.

For comparison purposes with other evaluation criteria, we construct the following summary measure of the overall goodness of fit: we average across the 7 measures of fit (4 for power spectra and 3 for coherences at business cycle frequencies) the difference between the median value across Monte Carlo replications and that expected if the model was the true DGP (0% in the first 4 cases and 100% in the last 3). The resulting values are:

Identification scheme:	Model 1:	Model 2:	Model 3:
Equal weights	3.87%	6.13%	70.29%
Different weights	16.97%	26.57%	73.71%

The first column can be associated to the "size" of the Watson (1993) evaluation methodology for calibrated models, whereas the 2nd and 3rd are measures of its "power". The table shows that, once the statistics of interest are selected, Watson's approach is substantially more accurate (better "size" and not much worse "power") when equal weights are given to all errors.

To summarize, extending Watson (1993) to evaluate calibrated models along multivariate dimensions is a reasonably accurate evaluation methodology. Not only the error associated by the Watson's measures of fit to testing the correct model is reasonable (small "size"), but also these measures are able to evaluate "how different" from the DGP alternative models are: the "power" is higher the more different the spectral density of the model is from that estimated for the actual data (in Figure 4.2 the spectral density matrix of Model 3 is more different from Model 1 than that of Model 2). Using a standard 5% significance level, Watson's approach (with an identification scheme of equal weights) would successfully accept only the true model (Model 1) and would also indicate that the error that should be added to Model 2 to match the DGP is less than a 10th of what Model 3 would need.

4.5 DeJong, Ingram and Whiteman's approach

DeJong, Ingram and Whiteman (DIW) (1996) propose a bayesian evaluation methodology for calibrated models which takes into account the uncertainty present in the statistics of both actual and simulated data to measure the fit of the model to the data. They suggest representing the actual data with a VAR and computing the distribution of the statistics of interest by drawing VAR parameters from their posterior distribution. In constructing distributions of simulated statistics, DIW consider only parameter uncertainty, and not the stochastic nature of the exogenous shocks as Canova and De Nicoló (1995). They use subjectively specified prior distributions (generally normal) for the parameters of the model whose location is set at the value typically calibrated in the literature while the dispersion is free. By enabling the specification of a sequence of increasingly diffuse priors over the parameter vector, the authors illustrate whether the uncertainty in the model's parameters can mitigate differences between the model and the actual data, so that the measure of dispersion can be used in order to (informally) minimize the distance between actual and simulated distributions of the statistics of interest.

As explained in Chapter 1, DIW suggest two statistics aimed at synthetically measuring the degree of overlap among actual and model statistics distributions. The first one is the Confidence Interval Criterion (CIC), which is defined as

$$CIC_{ij} = \frac{1}{1-\alpha}\int_a^b P_j(s_i)ds_i \tag{4.7}$$

where s_i, $i = 1, \ldots, n$, are the statistics of interest, $a = \frac{\alpha}{2}$ and $b = 1-a$ are the quantiles of $D(s_i)$, the distribution of the statistic in the actual data, $P_j(s_i)$ is the distribution of model statistic where j is the diffusion index of the prior on the parameter vector and $1 - \alpha = \int_a^b D(s_i)ds_i$. For CIC close to $\frac{1}{1-\alpha}$ the two distributions overlap substantially. If $CIC > 1$, $D(s_i)$ is diffuse relative to $P_j(s_i)$, i.e. the data is found to be relatively uninformative regarding s_i. For CIC close to zero, the fit of the model is poor, either because the overlap is small or because P_j is very diffuse. The second statistics DIW propose helps distinguishing among these two possible interpretations. The Difference of Means statistic is analogous to a t-statistic for the mean of $P_j(s_i)$ in the $D(s_i)$

distribution, i.e.

$$d_{ji} = \frac{EP_j(s_i) - ED(s_i)}{\sqrt{varD(s_i)}} \tag{4.8}$$

Large values of d_{ji} indicate that the location of $P_j(s_i)$ is quite different from the location of $D(s_i)$.

By providing a distribution rather than a single number, DIW methodology gives a more comprehensive characterization of actual and model statistics relative to the naive calibrator and also to Watson's approach, although these distributions rely on the subjective priors given by the researcher. The two measures of fit (CIC and d) are complementary and give a good summary of the fit of the model, which allows for comparison across models, e.g. the smaller is the average CIC across statistics of interest the better the fit.

The results of applying DIW methodology to our three models are summarized in Table 4.6. Prior distributions for parameters used for each model are shown in Table 4.5. DeJong, Ingram and Whiteman (1996) illustrate their methodology with the simpler version of the King, Plosser and Rebelo (1988) model, so our choice for the prior distributions is similar to theirs, although some parameters (especially the ones related to the exogenous government spending process) have been chosen using as reference Baxter and King (1993) and Aiyagari, Christiano and Eichenbaum (1992). We take 1000 draws of the parameter vector and compute at each draw the statistics std(C)/std(Y), corr(C,Y), corr(H,Y) and corr(H,AP) implied by the model[12]. For each model we characterize the statistics' distributions with the 5%, 50%, 95% percentiles, the mean and the standard deviation. Actual data statistics are computed fitting a VAR to linearly detrended logs of US data for 1964Q1-1995Q3 and randomizing its coefficients so that the statistics are computed for 1000 draws from the VAR coefficients distributions. The first lines in Table 4.6 summarize the distribution of actual data statistics.

We assess the fit of each model computing the percentage of the simulated statistics

[12]Instead of taking the theoretical values of the statistics, and for consistency with the statistical treatment of the actual data, we simulate time series for Y, C, H and AP 10,000 observations long and compute the statistics for their linearly detrended logs, so that these statistics are sufficiently close to their theoretical values.

laying between the 5% and 95% percentiles of the actual statistics distribution (CIC measure with $\alpha = 10\%$) and the standarized differences of means (d-statistic). The CIC and difference of means measures suggest a reasonable fit for Models 1 and 2 but, contrary to Watson's measure of fit, somehow better for the former with a higher overlap of distributions (average CIC of .87 versus .83 in Model 2) and with simulated statistics centered closer to actual ones in the case of Model 1 (smaller d-statistics). Model 1 statistics are less volatile than those observed for US data, while statistics from Model 2 and 3 are more volatile. This suggests that the standard deviation of government spending shocks has been left too volatile. Probably, reducing the standard deviation of its distribution would improve the fit of Model 2. The CIC and d-statistic for Model 3 clearly indicate a very bad fit. DeJong, Ingram and Whiteman (1996) obtain a worse fit than here for Model 1. There are two main reasons for this divergence. First, they evaluate 10 statistics of which std(C)/std(Y) and corr(C,Y) are the ones better captured by the model. Second, their actual data statistics differ from ours (different time period -1959Q1 to 1992Q2- and different detrending method -extracting a common time trend from consumption, investment and output-) especially those related to H. They measure H using average weekly hours of all workers instead of using per capita total hours (their measure of hours times employment divided by total population) as we do, following King, Plosser and Rebelo (1988)[13].

[13]This is an important difference, since we are including the evolution of both employment and hours in our measure of H whereas they only include that of hours worked by employees. A well known fact of the US economy is that about two thirds of the variation in total hours worked appears to be due to movements into and out of employment, while the remainder is due to adjustments in hours worked by employees. This fact has inspired a large number of real business cycle models which include nonconvexities in labor supply so that changes in total hours are brought by changing employment only (see Hansen (1985)) or changing both employment and hours per worker (see Cho and Cooley (1994)). The contemporaneous correlation between total hours and output reported in the literature for the US varies depending on the time period considered and the detrending method: using the Hodrick-Prescott filter Kydland and Prescott (1982) report a corr(H,Y) of 0.85, Hansen (1985) of 0.76, Cho and Cooley (1995) of 0.87 while King, Plosser and Rebelo (1988) report a contemporaneous corr(H,Y) of 0.07 extracting a common trend from output, consumption and investment. However, King, Plosser and Rebelo (1988) argue that this correlation rises considerably by splitting the sample into subperiods: they report that it averages 0.77 across subsamples. They interpret this sensitivity to the sample period as a suggestion that their detrending method may not have removed a low frequency component in output.

4.5.1 Evaluating DeJong, Ingram and Whiteman's approach

To evaluate the DIW methodology we conduct the same Monte Carlo experiments we have used before and test the following three hypotheses H_0: Model 1 = DGP, H_0: Model 2 = DGP and H_0: Model 3 = DGP. At each Monte Carlo replication we generate a distribution of model statistics using 100 draws from the corresponding prior distributions for the parameters, simulating at each draw long time series for Y, C, H and AP and computing the statistics of their linearly detrended logs. The distribution of actual DGP statistics is constructed similarly drawing from Model 1 priors but it is kept fixed across Monte Carlo replications and for testing all three hypothesis.

Table 4.7 displays the median and standard deviation (across Monte Carlo replications) of the 5%, 50% and 95% percentiles of the simulated distributions and, more importantly for evaluating the DIW methodology, medians and standard deviations of CIC (including the average CIC across statistics), *d*-statistic and the standarized difference of medians. For completeness, we have also computed the percentage of replications for which the difference of the medians exceeds 2 standard deviations of the actual statistic.

The overlap of DGP and Model 1 distributions is almost perfect (the median CIC across Monte Carlo replications is almost 1 in all 4 cases) and quite good but worse for Model 2, as expected. The *d*-statistic and the standarized difference of medians suggest that the worse fit of Model 2 is due to the fact that the mean and median of the DGP and model distributions are different, although the degree of overlap is high. This is especially the case for the corr(H,AP), as shown by the rejection frequency for the difference of medians (40% in Model 2 versus 1% in Model 1). That rejection frequency is 100% for all statistics but for std(C)/std(Y) under H_0: Model 3 = DGP. The methodology also reveals that among the four statistics, it is the relative standard deviation of C to Y the one that differs less between Models 3 and 1, both in the degree of overlap and in the location of the distributions. But the divergence is still clearly high.

Repeating what we have done with Watson's methodology, we construct a summary measure of the degree of rejection of a particular hypothesis, a measure which roughly captures the "size" and the "power" of the DIW methodology. For this purpose we

choose the average across the 4 statistics of the differences between the median CIC and their corresponding expected value if the model was the true DGP, i.e. equal to 1. It does not make much sense to include the difference between the *d*-statistic or standarized difference of medians and their expected values since they are measured in standard deviations of the actual mean or median, and hence not comparable to the divergence of the CIC measure to 1. The values of the summary measure are:

Model 1	Model 2	Model 3
1.25%	9%	90.5%

According to this ad-hoc summary measure, the DIW methodology seems to be much more accurate as a model evaluation methodology than Watson's, showing a smaller "size" (1.25%) and higher "power" especially against Model 3.

4.6 Canova and Canova and De Nicoló approaches

Canova and De Nicoló (1995) extend Canova (1994)-(1995) model evaluation methodology to a multivariate framework and compute measures of overlap of actual and model statistics in the same spirit as DIW but with some differences. Chapter 1 explains in detail these approaches and compares them.

Canova (1994)-(1995) takes the actual data statistics as fixed numbers and uses the uncertainty of simulated data to provide a measure of fit for the model. In addition to allowing the realization of the exogenous disturbance to vary, he also allows for parameter variability in measuring the dispersion of simulated statistics. As in DIW, parameters are considered uncertain not so much because of sample variability, but because there are many estimates of the same parameter obtained in the literature since estimation techniques, samples and frequency of the data tend to differ. Canova proposes to calibrate the parameter vector to an interval selected on the basis of these estimates, rather than to a particular value or than centering an arbitrarily diffuse prior normal to a particular value, as in DIW. Once the empirical distribution of the statistics of interest is constructed (simulating the model repeatedly by drawing parameter vectors from the postulated distribution and realizations of the exogenous stochastic process from some given distribution), one can then compute either the size

of calibration tests (using the actual statistics as a critical value for the simulated distribution) or the percentiles where the actual statistics lie.

Canova and De Nicoló (CDN) (1995), as DeJong, Ingram and Whiteman (1996), consider the uncertainty present in the statistics of both actual and model simulated data to measure the fit of the model to the data. CDN use a parametric bootstrap algorithm to construct distributions for the statistics of the actual data. In constructing distributions of simulated statistics, CDN take into account both the uncertainty in exogenous processes and parameters following Canova (1994)-(1995), while DIW only consider parameter uncertainty. To assess the degree of overlap of the two distributions, CDN choose a particular contour probability for one of the two distributions and ask how much of the other distribution is inside the contour. In other words, the fit of the model is examined very much in the style of the Monte Carlo literature: a good fit is indicated by a high probability covering of the two regions. To describe the features of the two distributions, they also repeat the exercise varying the chosen contour probability, say, from 50% to 75%, 90%, 95% and 99%. As in the DIW approach actual data and simulated data are used symmetrically, in the sense that in the CDN approach one can either ask whether the actual data could be generated by the model, or viceversa, whether simulated data are consistent with the distribution of the observed sample. As the DIW approach, CDN provides a more comprehensive evaluation of a calibrated model that Watson's or a naive calibrator approach.

We have evaluated our three models with the CDN approach in Tables 4.9 and 4.10. For consistency, we have computed separately the distributions of each statistic as in Canova (1994)-(1995) but consider distributions of both actual and model simulated statistics as in Canova and De Nicoló (1995). The distribution of actual data statistics has been constructed bootstrapping 1000 times the VAR fitted for the same US data used in previous sections (1964Q1-1995Q3). Statistics describing the bootstrap distribution of actual moments (5%, 50%, and 95% percentiles) are displayed in the top part of Table 4.9. The distribution is similar to that obtained using DIW (see Table 4.6) although std(C)/std(Y) is smaller and less volatile while corr(H,AP) is higher and more volatile. For each of the three models, we have simulated time series of the same sample size of actual data from the model 1000 times taking each time a draw

from the parameters' prior distributions described in Table 4.8 and a different random realization for the exogenous disturbances. At each draw we compute the 4 statistics for the simulated series, instead of taking their theoretical counterparts as in DIW. We are introducing two new sources of error with respect to DIW: a Monte Carlo error for the fact of taking random realizations of the shocks series and an estimation error for computing the simulated statistics from short time series for Y, C, H and AP. Parameter distributions are constructed just as the empirical based distributions used to evaluate Baxter and Crucini (1993) in Chapter 1. We have used existing estimates of these parameters in the literature or, when there are none, we have chosen a-priori an interval on the basis of theoretical considerations and imposed a uniform distribution on it. In comparison with the DIW approach, we drop the Normality assumption in many cases. Once the two distributions are constructed, we report the following measures of overlap: percentage of the distribution of simulated statistics into the 50%, 90% and 95% one-sided, 90% and 95% two-sided confidence intervals of the bootstrap distribution of actual statistics, and viceversa.

Model 1 statistics are smaller than actual US data ones (except for corr(H,AP)) but more volatile. They lay substantially inside the actual distributions but the distributions are not equally centered (a lower percentage of the simulated distributions lay inside the two-sided actual distributions) whereas the actual statistics, being higher in values, lay in higher percentage inside the two-sided confidence intervals of simulated statistics which include higher values than the one-sided confidence intervals. The introduction of government spending shocks makes simulated statistics of Model 2 even more volatile and their medians even lower than Model 1 ones (except for corr(H,Y)) so that the percentage of simulated statistics inside the two-sided confidence intervals of actual statistics gets smaller and the other way around for actual statistics into simulated distributions. In sum, the two distributions lay further apart while the coverage is still high overall. Model 3 statistics are less volatile than those of Model 1 or Model 2 (the volatility allowed in Table 4.8 for government spending shocks is smaller than that of technology shocks) but their median values lay so far from the actual ones that the two distributions hardly overlap when considering two-sided confidence intervals. For example, 100% of the actual corr(H,Y) are smaller than the median value of .99 implied

by Model 3, and 100% of simulated corr(C,Y) and corr(H,AP) are smaller than the actual ones (the median values implied by the model are -.94 and -.97, respectively).

Overall, the CDN methodology conducts an even more thorough analysis of the model and actual statistics distributions than the DIW approach. It discriminates better between the first two models, even though it gives a similar picture of the fit: Model 1 appears preferable to Model 2 as whith the DIW methodology. This is because Model 2 statistics are more volatile and centered further away from the actual ones. Nevertheless, the fit of Model 2 remains fair. As it happens with other model evaluation criteria, the CDN approach gives a bad fit to Model 3.

4.6.1 Evaluating Canova and De Nicoló approach

The results of the Monte Carlo experiment evaluating the performance of CDN are summarized in Table 4.11. As with the DIW approach, we keep the distribution of actual DGP moments fixed across Monte Carlo replications and for evaluating all three models. Such distribution has been generated simulating 100 times the DGP (Model 1) using a single realization of the technology shock process and one draw of the parameter vector from the prior distributions reported in column 1 of Table 4.8 at each iteration, and computing the statistics each time. At each Monte Carlo replication distributions of model statistics have been constructed by simulating 100 times the corresponding model, and one-sided and two-sided measures of overlap between actual and simulated distributions of each statistic have been computed. Table 4.11 displays their medians and standard deviations across the 100 Monte Carlo replications performed.

The last column of Table 4.11 presents the theoretical value of each measure of overlap which should be found if the model was the true DGP. The difference between the empirical values and the theoretical ones is an indication of the "size" of the CDN methodology when testing H_0: Model 1 = DGP, and of its "power" when testing H_0: Model 2 = DGP or H_0: Model 3 = DGP. Finally, we define a summary measure of the percentage rejection of each H_0 averaging across the 40 measures of overlap the difference between their median values across Monte Carlo replications and their expected true values. These summary measures are the following:

Model 1	Model 2	Model 3
0.55%	7.8%	53.4%

It is remarkable how accurately the CDN methodology recognises the true DGP: all 40 statistics presented in Table 4.11 are almost equal to their theoretical values. In fact the "size" measure is lower than using Watson or DIW methodologies. And this is particularly remarkable with respect to DIW, since as explained above we are introducing both a Monte Carlo and an estimation error when computing simulated statistics. This better "size" comes at one cost: the "power" against alternative models is in fact reduced. However, we still find that Model 3 is clearly rejected while the rejection of Model 2 is somewhat marginal.

4.7 Spectral density distance approach

The last approach we evaluate is the one presented in the previous chapter. We have derived an asymptotic test for the hypothesis that the quadratic standarized distance between the spectral density matrices of simulated and actual data is zero or less than an arbitrary prespecified bound. It is especially suitable for assessing the performance of models at a certain frequency range, such as business cycle models.

The test statistic proposed explicitly acknowledges that the solution paths generated by the model for the variables of interest are only approximations to the true model solution. Watson (1993) also recognises that there is an approximation error but, contrary to his approach, we take it into account to derive a formal test of the distance between the model and the observed data. Diebold, Ohanian and Berkowitz (1995) propose a measure of distance to evaluate how well the model matches the spectral density matrix of the actual data, too, but they assume that model spectra can be obtained without error. On the contrary, we compare actual to simulated data by treating them as samples from an unknown DGP and hence both spectral density matrices are estimated with error (the former because of sampling error, and the latter because of the approximation error). As in the DIW and CDN methodologies, the test proposed treats symmetrically actual and simulated data by taking into account the uncertainty existing in both data sets. While not excluding the possibility of stochastic

parameters in the model, the uncertainty considered in the model derives from the fact that there exists an approximation error. The main differences between this methodology and that of DIW and CDN are, first, that both sets of statistics are estimated in a classical instead of a Bayesian way and, second, that model and actual data are compared using asymptotic tests.

More specifically, the assessment of the fit of a model over a particular set of frequencies (e.g. business cycle frequencies $[\omega_1, \omega_2]$) is based on testing the following null hypothesis

$$H_0 : \Lambda D(\omega;\gamma) = 0 \quad , \quad \forall \omega \in [\omega_1, \omega_2]$$

where Λ is a selection matrix which weights the elements in the measure of distance $D(\omega;\gamma)$. $D(\omega;\gamma)$ is defined as

$$D(\omega;\gamma) = S\text{vec}f(\omega;\gamma) = \text{vec}f^{\mathbf{y}}(\omega) - \text{vec}f^{\mathbf{x}}(\omega;\gamma) \tag{4.9}$$

where $f(\omega;\gamma)$ is the spectral density matrix of the vector $[y_t \;\; x_t(\gamma, z_t)]$. y_t and x_t are the 1xN vectors of actual and simulated data, respectively. x_t depends on the model parameter vector γ and the exogenous shocks series z_t. $f^{\mathbf{y}}(\omega)$ and $f^{\mathbf{x}}(\omega;\gamma)$ are the upper left and lower right submatrices of $f(\omega;\gamma)$.

To test H_0, the following test statistic is proposed

$$fit([\omega_1,\omega_2];\gamma) = \sum_{\omega=\omega_1}^{\omega_2} \overline{\left(\sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma)\right)}' \left(\Lambda\Sigma_D(\omega;\gamma)\Lambda'\right)^{-1} \sqrt{\frac{\nu}{2}}\Lambda\hat{D}(\omega;\gamma) \tag{4.10}$$

where $\Sigma_D(\omega;\gamma)$ is the covariance matrix of $D(\omega;\gamma)$ and $\hat{D}(\omega;\gamma)$ is the estimated distance. The asymptotic distribution and properties of $\hat{D}(\omega;\gamma)$ are derived from those of the spectral density matrix estimator $\hat{f}(\omega;\gamma)$. Appropriately choosing the spectral window function and the bandwidth parameter (see Chapter 2) we can derive the following asymptotic distribution of the *fit* test statistic for each frequency under H_0

$$fit(\omega;\gamma) \;\dot{\sim}\; \chi^2_{(N^2-Q)} \quad , \omega \neq 0, \pm\pi \tag{4.11}$$

where Q is the number of zero elements in the diagonal of Λ. Therefore,

$$fit([\omega_1,\omega_2];\gamma) \dot{\sim} \chi^2_{L(N^2-Q)}$$

where L is the number of frequencies included in $[\omega_1, \omega_2]$.

The main advantage of this methodology relative to others is that it can reject or accept a model in a strict statistic sense, because such a statement is made by comparing a test statistic to a known asymptotic distribution. On the other hand, it provides only one measure of fit per frequency, while Watson's approach provides one per statistic (as the naive calibrator's), and CDN and DIW provide a wider variety of measures of fit.

Each time h the model is simulated, an $\hat{f}_h^{\mathbf{x}}(\omega;\gamma)$ is estimated keeping y_t fixed and using x_{ht}, for $h = 1,...,H$. In practice, what we are interested in obtaining is the average across the H replications of the estimated distance, i.e. $\hat{D}(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H}\hat{D}_h(\omega;\gamma) = \frac{1}{H}\sum_{h=1}^{H} S\ vec\hat{f}_h(\omega;\gamma)$. Given that x_{ht} are iid, that average keeps the same distribution and theoretical mean than $\hat{D}_h(\omega;\gamma)$, and $\Sigma_D(\omega;\gamma)$ becomes $\frac{1}{H}\Sigma_D(\omega;\gamma)$. The distribution for $fit([\omega_1,\omega_2];\gamma)$ keeps being valid as long as we premultiply $\hat{D}(\omega;\gamma)$ by $\sqrt{H}$ when constructing the test statistic. Then, H_0 will be rejected and the distance between the model and the actual data found significantly different from zero if $fit([\omega_1,\omega_2];\gamma)$ is greater than the critical value of a $\chi^2_{L(N^2-Q)}$, for a selected significance level α.

To assess the fit of our three models, we simulate 1000 times (H=1000) the model using the parameter vectors (γ) of Table 4.1 and compute $\hat{D}_h(\omega;\gamma)$ at each simulation by estimating the joint spectral density matrix for linearly detrended logs of the 4 actual US series and those simulated from the model. We have used a Quadratic Spectral window function and an optimal spectral window parameter estimate following Andrews (1991). Then we have taken the average $\hat{D}_h$ across simulations and computed $fit(\omega_1;\gamma)$, $fit(\omega_2;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$, where ω_1 (ω_2) are the frequencies associated with cycles 8 years (2 years) long. We have given equal weights to the elements in the spectral density submatrices $\hat{f}^{\mathbf{y}}(\omega)$ and $\hat{f}^{\mathbf{x}}(\omega;\gamma)$ (Q=0), therefore the asymptotic distributions of the three *fit* statistics are χ^2_{16}, χ^2_{16} and $\chi^2_{7\times16}$, respectively (the number of independent frequencies included in the $[\omega_1,\omega_2]$ interval depends on the value of the Andrews optimal bandwidth parameter which in turn depends on the parametric model fitted for the actual data, and is 7 in this case). To evaluate the $fit([\omega_1,\omega_2];\gamma)$ statistic, we have used the following Normal approximation typically used for χ^2_k distributions of k > 100 : $\sqrt{2\,\chi^2_k} \sim N(\sqrt{2k-1};1)$.

Following we report the *fit* statistics and the 90% and 95% critical values:

	Model 1	Model 2	Model 3	90% C.V.	95% C.V.
$fit(\omega_1;\gamma)$	7379	7614	29819	23.5	26.3
$fit(\omega_2;\gamma)$	7264	7537	29342	23.5	26.3
$fit([\omega_1,\omega_2];\gamma)$	11224	11588	45863	137.3	142.7

It turns out that none of the models is accepted as the US data DGP: the values of all test statistics are clearly grater than the critical values in all cases. As pointed out in Chapter 2, the sample size of the data used is too short to assume the asymptotic normality of $\hat{D}(\omega;\gamma)$. Its distribution may be closer to a χ^2 and therefore how many $\hat{D}_h(\omega;\gamma)$ it aggregates matters, so that eventually the asymptotic distribution of $fit(\omega;\gamma)$ may converge to a χ^2 distribution too, but with degrees of freedom increasing with H (and hence, the critical values too).

On the other hand, and consistently to other model evaluation methodologies, Model 1 and Model 2 are found to be almost equally closer to the actual data and Model 3 four times more distant. Contrary to the CDN and DIW methodologies (both allowing for random parameters in the models) and just as in the Watson methodology (which uses the parameters in Table 4.1, too), Model 1 appears closer to the US data than Model 2.

4.7.1 Evaluating the spectral density distance approach

We finally perform the Monte Carlo experiment on this last methodology. At each replication, one realization of the 4 time series from the DGP (using Model 1 with fixed parameters) is compared to one of the 4 simulated series from the corresponding model, for each of the 100 times the model is simulated, and we compute the average estimated distance across the 100 simulations to calculate the $fit(\omega_1;\gamma)$, $fit(\omega_2;\gamma)$ and $fit([\omega_1,\omega_2];\gamma)$ statistics as explained above.

Table 4.12 summarizes the performance of the *fit* test statistics in testing H_0: Model i = DGP, for $i = 1, 2$ and 3. It displays the percentage rejection of each hypothesis when comparing the corresponding *fit* test statistic to its 90% and 95% critical values, and the 5%, 50%, 90%, 95% percentiles and the mean and standard deviation of each of the

three *fit* statistics computed across the 100 replications of the Monte Carlo experiment. Now the number of frequencies included in the business cycle interval is 15 instead of 7 (because the optimal bandwidth parameter has changed since the actual data now is not the US observed data but that simulated from the DGP -Model 1-) and hence the critical values for the $fit([\omega_1, \omega_2]; \gamma)$ test statistic change. Figure 4.3 plots the average across Monte Carlo replications of the *fit* statistic for each frequency.

The first two rows of statistics reported in Table 4.12 for testing each hypothesis are the empirical size (for H_0: Model 1 = DGP) and power (for H_0: Model 2 = DGP and H_0: Model 3 = DGP) of the spectral density distance methodology. A summary table comparable to the measures of the size and power we have computed for the other methodologies would be:

	Model 1	Model 2	Model 3
significance level 5%	0%	0%	100%
significance level 10%	0%	0%	100%

The small size found (0% versus theoretical 5% or 10%), is consistent with the Monte Carlo experiment on the small sample properties of the *fit* test statistic performed in Chapter 2.

However, it seems that the power against models not too different from the DGP (Model 2) is null. The actual values of the *fit* statistic clearly indicate a worse fit for Model 2 than for Model 1 (as should be the case) but not bad enough to reject that the spectral density of Model 2 is equal to that of the DGP, contrary to the previous methodologies. Part of the reason can be in that the *fit* test is a single overall measure of fit, while Watson's approach and especially DIW and CDN approaches can capture the discrepancy between the model and the DGP along many dimensions (they compute several measures of fit for each statistic), and this property is kept even when aggregating their different measures of fit into a summary one as we present at the end of each section. However, recall that this methodology tests the distance between model and actual spectral densities. Figures 4.1 and 4.2 clearly show that the frequency domain properties of Model 1 (our DGP) and Model 2 are almost indistinguishable. When models, as it is the case for most dynamic stochastic general equilibrium models, are known to be false, we want the evaluation methodology to be able to capture how

well they reproduce certain particular statistics. It turns out that two models different but close to each other as Model 2 to Model 1 may generate almost the same statistics we are interested in, in our case the multivariate spectral density matrix for Y, C, H and AP. We interpret the apparent inability for discriminating between Model 1 and Model 2 shown when evaluating both of them as exactly equal to the DGP (0% rejection) as the correct indication that both reproduce almost exactly the precise statistic of the DGP we want them to replicate. Hence we should be equally happy with both of them just as we would be equally unhappy if the DGP's spectral density differs equally from both. In fact, consistently with the results in Chapter 2, the *fit* test is very powerful when the alternative hypothesis imply different spectral density matrices to the DGP (see Model 3 spectral properties in Figure 4.2).

The issue arising here is an important one: before using any model evaluation methodology, it should be checked whether discriminating models according to the statistics they focus on (in this case, the spectral densities at particular frequencies) is desirable. It is highly probable that the spectral densities at low frequencies (such as business cycle frequencies) of series with similar autorregressive structures with high persistence parameters will not significantly differ, as was shown in Chapter 2. Many real business cycle model series follow that structure. On top of that, using a detrending method which does not totally remove very low frequency movements (as the linear detending method) concentrates relatively less spectral density at other frequencies, making harder to discriminate models according to their spectra at, say, the higher frequencies included in the business cycle range.

Probably, the real business cycle models we have chosen for performing our comparison exercise yield more observable differences between alternative models when evaluated using time domain statistics such as relative standard deviations and correlations than when using frequency domain statistics. A simple look at the discrepancies observed between the statistics simulated from Model 1 and 2 in Table 4.2 as compared to the difference between spectra and coherencies simulated from the same two models in Figures 4.1 and 4.2 confirms this point.

4.8 Conclusions

In this chapter we compare under uniform conditions the performance of alternative methodologies recently proposed in the literature to evaluate dynamic stochastic general equilibrium models.

We have first described the approaches, emphasizing the differences among them and with the standard informal evaluation approach. Second, we have illustrated the methodologies of Watson (1993), DeJong, Ingram and Whiteman (1996), Canova and De Nicoló (1995) and the one based on spectral density distance presented in Chapter 2, using three versions of a simple one-sector real business cycle model from King, Plosser and Rebelo (1988). Government shocks seem to add little or none explanatory power to technology shocks in a one-sector dynamic stochastic general equilibrium model for the US, and are certainly not enough to provide a reasonable fit when considered as the only source of fluctuations.

The main contribution of this chapter is to conduct a Monte Carlo experiment on the four methodologies to "test" them and compare their performance as evaluation procedures for dynamic general equilibrium models. We have encountered several difficulties in undertaking this task, which are mainly related to four facts. First, the comparison is made on a multivariate level, which complicates the effort of summarizing the overall performance of each methodology. Second, some methodologies are constructed in the frequency domain (Watson's and the spectral density distance approaches) while others are built in the time domain (DIW and CDN). Third, DIW and CDN define distributions for the parameters of the model in different ways while the two other approaches take parameters as fixed. Fourth, Watson and DIW use the theoretical values of model statistics (DIW also for actual data statistics) while CDN and the spectral density distance approach estimate them. Despite of these difficulties, we have been able to compute rough measures of the "size" and "power" of each model evaluation methodology.

Our exercise highlights that there are differences between the methodologies along many dimensions, but looking at the summary comparison provided by the "size" and "power" measures it is the two approaches allowing for stochastic parameters (DIW and CDN) the ones that seem to achieve a better performance. Probably, the real

business cycle models we have chosen for performing our comparison exercise yield more observable differences between models when evaluated using time domain statistics such as relative standard deviations and correlations than when using frequency domain statistics. A simple look at the discrepancies observed between the statistics simulated from Model 1 and 2 in Table 4.2 as compared to the difference between spectra and coherencies simulated from the same two models in Figures 4.1 and 4.2 confirms this point.

In fact, although the Spectral Density Distance approach presented in Chapter 2 is the one which obtains the smaller size and the larger power against models very different to the DGP (Model 3), it shows no power against Model 2. The spectral density matrices of Model 1 and 2 do not differ enough for the methodology to recognise them as different models. Watson's approach, also in the frequency domain, has significantly worse size and worse power against Model 3, but captures some discrepancy between the spectral properties of Model 1 and 2.

The time domain approaches of DeJong, Ingram and Whiteman (1996) and Canova and De Nicoló (1995) appear more accurate than Watson's. Among this two approaches, CDN achieves a better "size" at the cost of a lower "power", which is still enough to correctly rank the models according to their discrepancy with the true DGP.

We find that all four methodologies outperform the naive calibrator's rule since they substantially reduce the risk of rejecting the true DGP, are able to discriminate more clearly between the DGP and models very distant from it and all but the spectral density distance approach (for the reasons explained above) also have power against models whose DGP is slightly different to the true DGP.

Table 4.1: **Baseline parameter values**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor in Output (α)	0.58	0.58	0.58
Growth rate (θ_x)	1.0036	1.0036	1.0036
Depreciation Rate of Capital (δ_K)	0.025	0.025	0.025
Discount Factor (β)	0.9875	0.9875	0.9875
Steady State hours ($\bar{H}$)	0.20	0.20	0.20
Risk Aversion (σ)	2	2	2
Share of Government Spending in Output (sg)	0.25	0.25	0.25
Tax Rate (τ)	0.25	0.25	0.25
Persistence of Technology Disturbances (ρ_A)	0.9	0.9	0
Persistence of Government Spending Disturbances (ρ_G)	0	0.97	0.97
Standard Deviation of Technology Innovations (σ_A)	0.00852	0.00852	0
Standard Deviation of Government Spending Innovations (σ_G)	0	0.0036	0.0036

Table 4.2: **Actual data and simulated moments**

Statistic	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks	Actual Data
std(C)/std(Y)	.667 (.088)	.671 (.094)	.717 (.006)	.826
corr(C,Y)	.869 (.038)	.859 (.044)	-.999 (.0003)	.863 (.133)
corr(H,Y)	.776 (.097)	.765 (.108)	.999 (0)	.807 (.157)
corr(H,AP)	.374 (.171)	.344 (.182)	-.999 (.0001)	-.065 (.265)

Notes: Moments of both model simulated series and actual data are computed after linearly detrending the series.

Actual data are logs of per capita real variables in $Mln and for the period 1964Q1-1995Q3. Newey and West (1987) consistent S.E. are reported for the correlation coefficients.

Simulated statistics are average (std) across 100 simulations of the corresponding model where at each simulation different random series are used for the exogenous shocks and time series for Y, C, H and AP are generated of sample size equal to the actual data (127 observations). The random number generator is seeded at 0 before simulating each model. When persistence parameters or S.D. of innovations are 0, model simulations are run using $1x10^{-10}$ instead to avoid non-full rank matrix problems.

Table 4.3: **Watson's measures of fit. Averages across BC frequencies**

	sp(Y)	sp(C)	sp(H)	sp(AP)	cohe(C,Y)	cohe(H,Y)	cohe(H,AP)
Actual data statistics	.0004	.0003	.0002	.0001	.79	.64	.04
Model 1 statistics	.0004	.0001	.0001	.0001	.83	.87	.52
	Measures of Fit for Model 1 (only A_t shocks)						
	.027	.22	.18	.31	1.05	1.35	13.12
	.012	.46	.50	1	1.05	1.35	13.12
Model 2 statistics	.0004	.0001	.0001	.0001	.80	.85	.46
	Measures of Fit for Model 2 (both A_t and G_t shocks)						
Equal Weight	.026	.22	.16	.28	1.02	1.31	11.64
Min errors(Y,H)	.13	.82	.26	1.61	1.02	1.31	11.64
	.00002	.00001	.00004	.00001	.997	.999	.998
	Measures of Fit for Model 3 (only G_t shocks)						
Equal Weight	.91	1.05	.74	.97	1.27	1.55	25.21
Min error(Y)	.87	1.11	.78	1.10	1.27	1.55	25.21

See text for explanation

Table 4.4: **Monte Carlo on Watson's measures of fit**

Summary statistics	sp(Y)	sp(C)	sp(H)	sp(AP)	cohe(C,Y)	cohe(H,Y)	cohe(H,AP)
Testing H_0: Model 1 = DGP							
Identification: equal weights							
mean	.07	.15	.09	.15	1.005	1.05	1.16
std	.10	.23	.09	.23	.07	.15	.53
5%perc	.008	.023	.024	.023	.90	.91	.67
median	.034	.074	.062	.074	.99	1.007	.99
95%perc	.24	.55	.27	.55	1.14	1.34	2.13
Identification: min error(Y)							
mean	.065	.48	.44	.48	1.005	1.05	1.16
std	.10	.30	.20	.30	.07	.15	.53
5%perc	.006	.26	.22	.26	.90	.91	.67
median	.031	.38	.37	.38	.99	1.007	.99
95%perc	.24	1.04	.82	1.04	1.14	1.34	2.13
Testing H_1: Model 2 = DGP							
Identification: equal weights							
mean	.07	.15	.10	.15	.97	1.03	1.03
std	.10	.24	.11	.24	.07	.15	.47
5%perc	.009	.023	.027	.023	.87	.89	.59
median	.035	.073	.068	.073	.96	.98	.88
95%perc	.25	.57	.32	.57	1.11	1.31	1.89
Identification: min error(Y) and error(H)							
mean	.12	.79	.25	.79	.97	1.03	1.03
std	.11	.37	.15	.37	.07	.15	.47
5%perc	.05	.46	.12	.46	.87	.89	.59
median	.09	.69	.21	.69	.96	.98	.88
95%perc	.31	1.48	.53	1.48	1.11	1.31	1.89
Testing H_2: Model 3 = DGP							
Identification: equal weights							
mean	.90	1.06	.63	1.06	1.21	1.22	2.22
std	.03	.06	.05	.06	.09	.18	1.02
5%perc	.85	.98	.54	.98	1.09	1.05	1.28
median	.90	1.05	.64	1.05	1.20	1.16	1.92
95%perc	.95	1.17	.72	1.17	1.38	1.55	4.10
Identification: min error(Y)							
mean	.86	1.17	.70	1.17	1.21	1.22	2.22
std	.02	.05	.06	.05	.09	.18	1.02
5%perc	.81	1.10	.59	1.10	1.09	1.05	1.28
median	.86	1.16	.70	1.16	1.20	1.16	1.92
95%perc	.90	1.28	.79	1.28	1.38	1.55	4.10

Empirical distribution of Watson's Measures of Fit at business cycle frequencies over 1000 replications. Measures of Fit at each replication are computed as in Table 4.3.

Table 4.5: **Parameter distributions for the DIW methodology**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor in Output (α)	N(0.58,0.05)	N(0.58,0.05)	N(0.58,0.05)
Growth rate (θ_x)	1.0036	1.0036	1.0036
Depreciation Rate of Capital (δ_K)	N(0.025,0.004)	N(0.025,0.004)	N(0.025,0.004)
Discount Factor (β)	N(0.988,0.001)	N(0.988,0.001)	N(0.988,0.001)
Steady State hours ($\bar{H}$)	N(0.20,0.02)	N(0.20,0.02)	N(0.20,0.02)
Risk Aversion (σ)	N(2,1)	N(2,1)	N(2,1)
Share of Government Spending in Output (sg)	N(0.25,0.05)	N(0.25,0.05)	N(0.25,0.05)
Persistence of Technology Disturbances (ρ_A)	N(0.9,0.25)	N(0.9,0.25)	0
Persistence of Government Spending Disturbances (ρ_G)	0	N(0.97,0.02)	N(0.97,0.02)
Standard Deviation of Technology Innovations (σ_A)	N(0.00852,0.004)	N(0.00852,0.004)	0
Standard Deviation of Government Spending Innovations (σ_G)	0	N(0.0036,0.002)	N(0.0036,0.002)

Table 4.6: **DeJong, Ingram and Whiteman methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
US Data, 1964Q1-1995Q3				
5% perc	.72	.77	.82	-.16
median	.85	.87	.89	.12
95% perc	1.11	.96	.96	.60
mean	.89	.81	.85	.09
std	.46	.29	.16	.39
Simulated statistics, Model 1				
5% perc	.71	.78	.004	-.40
median	.87	.89	.50	.04
95% perc	1.07	.95	.73	.33
mean	.88	.88	.45	.01
std	.11	.06	.23	.22
Evaluating Model 1				
CIC	1.07	1.07	0.26	1.09
Average(CIC)	.87			
d-statistic	-.03	.33	-2.48	-.23
Simulated statistics, Model 2				
5% perc	.67	.07	.13	-.01
median	.87	.87	.53	-.15
95% perc	1.25	.95	.99	-.45
mean	1.41	.79	.51	-.03
std	3.07	.23	.26	.23
Evaluating Model 2				
CIC	.99	.98	.27	1.09
Average(CIC)	.83			
d-statistic	1.12	-.10	-2.13	-.33
Simulated statistics, Model 3				
5% perc	.20	-.20	.76	.10
median	.74	-.90	.99	-.97
95% perc	2.02	-.99	1	-.99
mean	1.36	-.75	.95	-.84
std	3.11	.35	.11	.37
Evaluating Model 3				
CIC	.45	.006	.22	.097
Average(CIC)	.19			
d-statistic	1.03	-7.85	.63	-2.38

Notes: Actual data statistics are computed fitting a VAR to linearly detrended logs of US data for 1964Q1-1995Q3 and randomizing its coefficients so that the statistics are computed for 1000 draws from the VAR coefficients distributions.
Simulated statistics are computed for series of 10,000 observations simulated from each model 1000 times, using at each simulation a different draw from the prior distributions of the parameters in Table 4.5.

Table 4.7: Monte Carlo on DIW methodology

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
Testing H_0: Model 1 = DGP				
Simulated statistics, Model 1				
5% perc	.71 (.02)	.77 (.04)	.02 (.11)	-.41 (.08)
median	.87 (.01)	.89 (.006)	.50 (.02)	.03 (.03)
95% perc	1.06 (.03)	.95 (.007)	.72 (.02)	.30 (.07)
Evaluating Model 1				
CIC	1.01 (.03)	.99 (.04)	1.01 (.03)	1.02 (.03)
Average(CIC)	1.003 (.02)			
d-statistic	.001 (.09)	.0009 (.11)	.071 (.09)	-.058 (.08)
Diff.Medians	-.04 (.11)	-.06 (.09)	.02 (.10)	-.05 (.13)
Rej. freq of Diff.Medians	0%	0%	0%	1%
Testing H_0: Model 2 = DGP				
Simulated statistics, Model 2				
5% perc	.65 (.03)	.12 (.25)	.14 (.10)	.001 (.015)
median	.86 (.01)	.86 (.009)	.54 (.02)	.13 (.15)
95% perc	1.15 (2.69)	.94 (.008)	.99 (.06)	.32 (.46)
Evaluating Model 2				
CIC	.91 (.05)	.82 (.04)	.93 (.04)	.98 (.04)
Average(CIC)	.91 (.03)			
d-statistic	-.004 (.11)	-.54 (.27)	.28 (.11)	.25 (.13)
Diff.Medians	-.10 (.12)	-.42 (.14)	.16 (.09)	.52 (.66)
Rej. freq of Diff.Medians	0%	0%	0%	40%
Testing H_0: Model 3 = DGP				
Simulated statistics, Model 3				
5% perc	.18 (.07)	.07 (.18)	.72 (.09)	.024 (.06)
median	.73 (.06)	-.90 (.02)	.99 (.002)	-.97 (.006)
95% perc	1.89 (2.55)	-.999 (.0003)	1 (0)	-.999 (.001)
Evaluating Model 3				
CIC	.25 (.05)	0 (.003)	.06 (.02)	.07 (.02)
Average(CIC)	.097 (.014)			
d-statistic	.64 (.15)	-19 (.17)	2.14 (.05)	-5 (.04)
Diff.Medians	-1.25 (.56)	-27.36 (.36)	2.11 (.008)	-4.52 (.03)
Rej. freq of Diff.Medians	54%	100%	100%	100%

Medians (standard deviations) across 100 Monte Carlo replications of summary statistics of the simulated distributions and of DIW model evaluation statistics (CIC and *d*-statistic) and related. See text.

Table 4.8: **Parameter distributions for the CDN methodology**

Parameter	Model 1: Only A_t shocks	Model 2: A_t and G_t shocks	Model 3: Only G_t shocks
Share of Labor (α)	U[0.5,0.75]	U[0.5,0.75]	U[0.5,0.75]
Growth rate (θ_x)	N(1.0036,0.001)	N(1.0036,0.001)	N(1.0036,0.001)
Depreciation Rate of Capital (δ_K)	U[0.02,0.03]	U[0.02,0.03]	U[0.02,0.03]
Discount Factor (β)	TruncN[0.9855,1.002]	TruncN[0.9855,1.002]	TruncN[0.9855,1.002]
St.St. Hours ($\bar{H}$)	U[0.2,0.35]	U[0.2,0.35]	U[0.2,0.35]
Risk Aversion (σ)	Truncχ^2(2)[0,10]	Truncχ^2(2)[0,10]	Truncχ^2(2)[0,10]
Share of G (sg)	U[0.2,0.3]	U[0.2,0.3]	U[0.2,0.3]
Persistence of Tech. Disturbances (ρ_A)	N(0.9,0.2)	N(0.9,0.2)	0
Persistence of G Disturbances (ρ_G)	0	U[0.95,0.9999]	U[0.95,0.9999]
Std of Technology Innovations (σ_A)	Truncχ^2(1)[0,0.0202]	Truncχ^2(1)[0,0.0202]	0
Std of G Innovations (σ_G)	0	Truncχ^2(1)[0,0.01]	Truncχ^2(1)[0,0.01]

Table 4.9: **Canova and De Nicoló methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
US Data				
5% perc	.60	.72	.83	-.33
median	.76	.89	.93	.15
95% perc	.98	.96	.97	.57
Simulated statistics, Model 1				
5% perc	.47	.48	.47	-.38
median	.65	.82	.79	.29
95% perc	.91	.94	.95	.72
Evaluating Model 1				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	77	77	94	35
90% one-sided C.I.	96	96	99	75
95% one-sided C.I.	98	98	99	83
90% two-sided C.I.	63	78	40	76
95% two-sided C.I.	74	84	47	85
% of actual statistics into simulated distributions'				
50% one-sided C.I.	13	22	2	69
90% one-sided C.I.	78	74	35	98
95% one-sided C.I.	89	86	55	99
90% two-sided C.I.	89	85	55	96
95% two-sided C.I.	93	93	65	98

Table 4.10: **Canova and De Nicoló methodology (cont.)**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)
Simulated statistics, Model 2				
5% perc	.39	-.87	.50	-.94
median	.63	.76	.82	.11
95% perc	.93	.93	.98	.67
Evaluating Model 2				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	78	84	84	53
90% one-sided C.I.	95	98	91	82
95% one-sided C.I.	97	99	92	87
90% two-sided C.I.	55	56	40	65
95% two-sided C.I.	66	62	47	71
% of actual statistics into simulated distributions'				
50% one-sided C.I.	9	8	4	44
90% one-sided C.I.	78	61	83	96
95% one-sided C.I.	91	79	99	98
90% two-sided C.I.	90	79	99	98
95% two-sided C.I.	95	87	100	99
Simulated statistics, Model 3				
5% perc	.28	-.99	.96	-.99
median	.54	-.94	.99	-.98
95% perc	.96	-.71	1	-.86
Evaluating Model 3				
% of simulated statistics into actual distributions'				
50% one-sided C.I.	81	100	2	100
90% one-sided C.I.	92	100	6	100
95% one-sided C.I.	96	100	8	100
90% two-sided C.I.	37	0	8	0
95% two-sided C.I.	45	0	12	0
% of actual statistics into simulated distributions'				
50% one-sided C.I.	1	0	100	0
90% one-sided C.I.	84	0	100	0
95% one-sided C.I.	94	0	100	0.1
90% two-sided C.I.	94	0	14	0.1
95% two-sided C.I.	96	0	31	0.1

Table 4.11: **Monte Carlo on the CDN methodology**

	std(C)/std(Y)	corr(C,Y)	corr(H,Y)	corr(H,AP)	
Testing H_0: Model 1 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	51 (5.1)	50 (5)	50 (5.1)	50 (5)	50
90% one-sided C.I.	90 (3.1)	91 (2.8)	90 (3.1)	90 (3)	90
95% one-sided C.I.	94 (2.4)	96 (2)	95 (2)	95 (2)	95
90% two-sided C.I.	88 (3.3)	91 (2.8)	90 (3)	90 (3)	90
95% two-sided C.I.	94 (2.5)	96 (2)	94 (2)	95 (2.2)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	49 (5.4)	50 (5)	50.5(5.2)	50 (4.8)	50
90% one-sided C.I.	90 (3.7)	88.5(3.3)	90.5(2.7)	90 (3)	90
95% one-sided C.I.	96 (2.5)	95 (2.3)	95 (2.1)	94 (2)	95
90% two-sided C.I.	91 (3)	88 (3)	90 (3)	89 (3)	90
95% two-sided C.I.	95 (2.1)	94 (2.6)	95 (2.2)	94 (2.3)	95
Testing H_0: Model 2 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	55 (5)	66 (4.8)	42 (4.9)	64 (4.7)	50
90% one-sided C.I.	89 (3.1)	94 (2.2)	79 (3.9)	93 (2.6)	90
95% one-sided C.I.	93 (2.5)	97 (1.7)	85 (3.5)	97 (1.7)	95
90% two-sided C.I.	81 (3.9)	73 (4.5)	81 (3.9)	75 (4.3)	90
95% two-sided C.I.	88 (3.2)	79 (4.1)	85 (3.6)	80 (3.9)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	44 (6)	28 (5)	58 (5)	33 (5.2)	50
90% one-sided C.I.	91 (4)	83 (3.3)	99 (1.7)	85 (4)	90
95% one-sided C.I.	97 (2)	92(2.3)	100(.4)	93 (3)	95
90% two-sided C.I.	95.5(2.4)	92 (3)	93.4(3.2)	93 (3.1)	90
95% two-sided C.I.	98 (1.3)	95.5(2.6)	97.3(2.2)	95.5(2.2)	95
Testing H_0: Model 3 = DGP					
% of simulated statistics into actual distributions'					
50% one-sided C.I.	66.5(5)	100 (0)	0 (.3)	100 (0)	50
90% one-sided C.I.	89 (3)	100 (0)	1 (.9)	100 (0)	90
95% one-sided C.I.	93 (2.6)	100 (0)	1 (1.2)	100 (0)	95
90% two-sided C.I.	55 (5.1)	0 (0)	1 (1.2)	0 (.1)	90
95% two-sided C.I.	63 (5)	0 (.04)	2 (1.5)	0 (.2)	95
% of actual statistics into simulated distributions'					
50% one-sided C.I.	17 (7.6)	0 (0)	100 (0)	0 (0)	50
90% one-sided C.I.	92 (4)	0 (0)	100 (0)	0 (0)	90
95% one-sided C.I.	97 (2)	0 (0)	100 (0)	0 (.01)	95
90% two-sided C.I.	97 (2)	0 (0)	0.9(0.9)	0 (.01)	90
95% two-sided C.I.	98.7(1.2)	0 (0)	2 (4)	0 (.08)	95

dians (S.D.) across 100 Monte Carlo replications of the CDN measures of percentage overlap ween the distributions of actual and model statistics.

Table 4.12: **Monte Carlo on the spectral density distance methodology**

	$fit(\omega_1;\gamma)$	$fit(\omega_2;\gamma)$	$fit([\omega_1,\omega_2];\gamma)$
Testing H_0: Model 1 = DGP			
% rejection (90% C.I.)	0%	0%	0%
% rejection (95% C.I.)	0%	0%	0%
5% perc	1.55	1.47	31.45
median	3.42	3.29	52.8
90% perc	7.19	5.3	73.79
95% perc	7.56	5.43	91.72
mean	3.84	3.50	53.61
S.D.	1.99	1.48	17.73
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6
Testing H_0: Model 2 = DGP			
% rejection (90% C.I.)	0%	0%	0%
% rejection (95% C.I.)	0%	0%	0%
5% perc	2.23	2.67	49.09
median	5.8	5.16	76.7
90% perc	9.1	7.4	117.5
95% perc	10.6	8.13	123.78
mean	6.09	5.18	81.26
S.D.	2.76	1.96	26.6
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6
Testing H_0: Model 3 = DGP			
% rejection (90% C.I.)	100%	100%	100%
% rejection (95% C.I.)	100%	100%	100%
5% perc	338.3	407.4	5721.1
median	351.5	418.3	5836.2
90% perc	364	428	5908
95% perc	367.7	430	5921
mean	352	418.6	5834
S.D.	9.84	7.17	61.08
90% C.V	23.5	23.5	172.63
95% C.V.	26.3	26.3	178.6

Summary statistics of the empirical distribution across 100 Monte Carlo replications of the *fit* test statistics for frequencies associated to cycles 8 years long (ω_1), 2 years long (ω_2) and for averages across business cycle frequencies ($[\omega_1,\omega_2]$ interval).

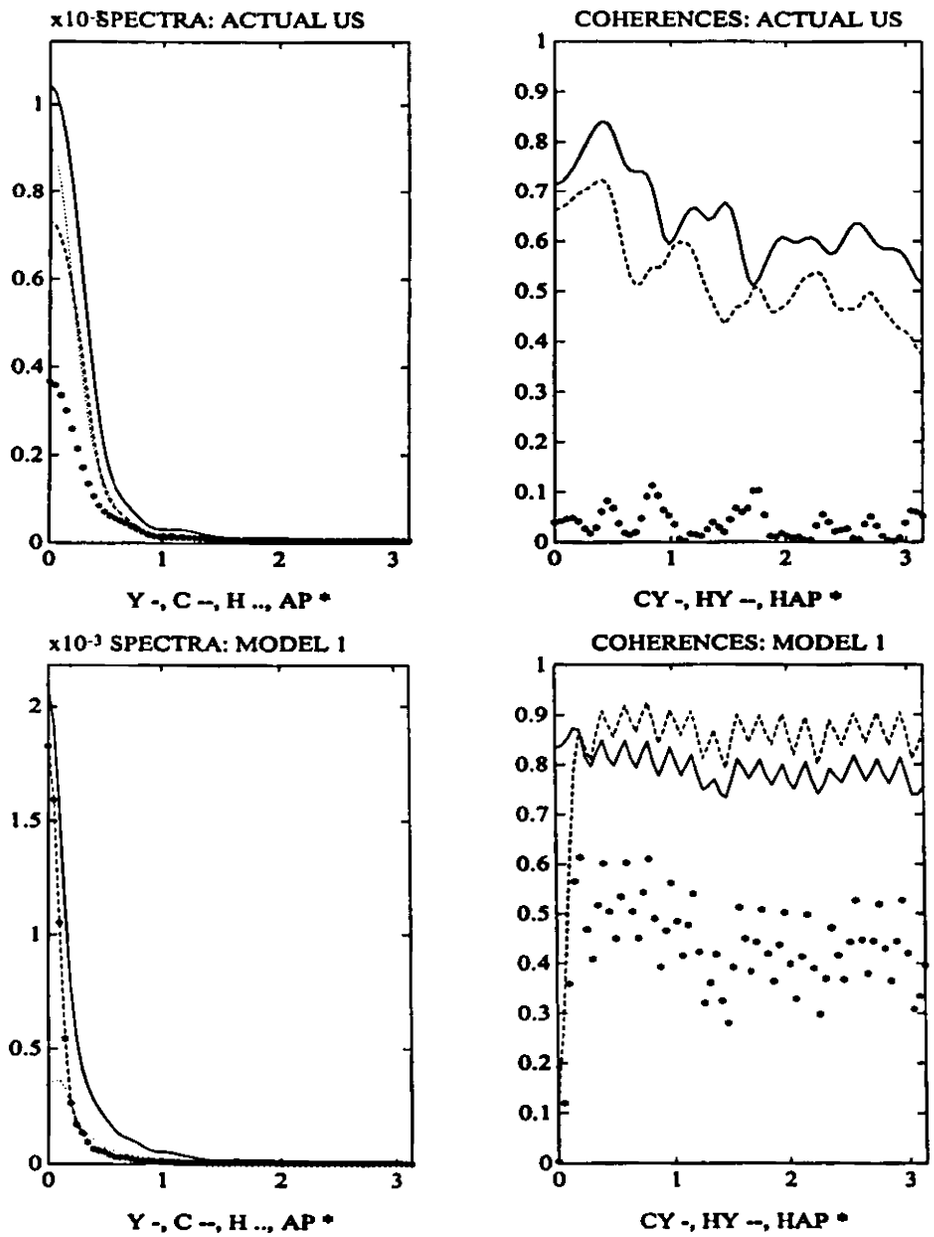

Figure 4.1: Spectra of Y(–),C(- -),H(..) and AP(*) and coherences of C,Y(–), H,Y(- -) and H,AP(*). **Actual US data** in the upper plots, simulated data from **Model 1** in the lower plots. A linear trend has been extracted from all series.

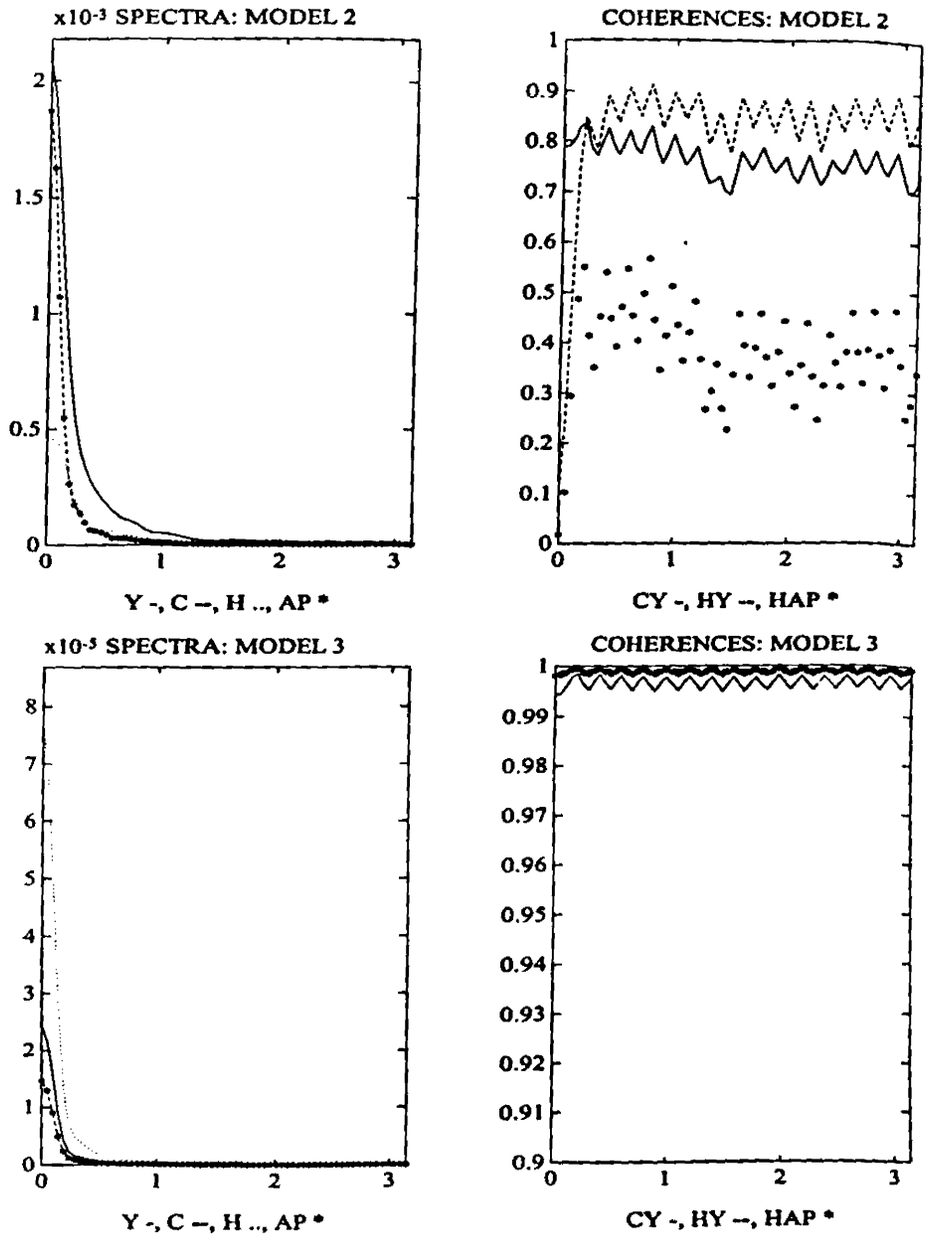

Figure 4.2: Spectra of Y(–),C(- -),H(..) and AP(*) and coherences of C,Y(–), H,Y(- -) and H,AP(*). Simulated data from **Model 2** in the upper plots, simulated data from **Model 3** in the lower plots. A linear trend has been extracted from all series.

Bibliography

[1] Aiyagari, R., Christiano, L. and M. Eichenbaum (1992) "The Output, Employment, and Interest Rate Effects of Government Consumption", *Journal of Monetary Economics*, 30, 73-86.

[2] Andersen, T.G. and B.E. Sorensen (1995), "GMM Estimation of a Stochastic Volatility Model: a Monte Carlo Study", University of Copenhagen Discussion Paper 95-19.

[3] Andrews, D.W. (1991), "Heteroskedasticity and autocorrelation consistent covariance matrix estimation", *Econometrica*, vol.59, No.3, 817-858.

[4] Armington, P. (1996), "A Theory of Demand for Products Distinguished by Place of Production", International Monetary Fund Staff Papers, 27, 159-178.

[5] Backus, D., Kehoe, P. and F. Kydland (1992), "International Real Business Cycles", *Journal of Political Economy*, vol.100, 745-775.

[6] Backus, D., Kehoe, P. and F. Kydland (1993), "Dynamics of Trade Balance and the Terms of Trade: the J-curve?", *American Economic Review*, 84, 84-103.

[7] Backus, D., Kehoe, P. and F. Kydland (1995), "International Business Cycles: Theory and Evidence", in T. Cooley (ed.), *Frontiers of Business Cycle Analysis*, Princeton, NJ: Princeton University Press.

[8] Baxter, M. (1991) "Approximating Suboptimal Dynamic Equilibria: An Euler Equation Approach", *Journal of Monetary Economics*, 27, 173-200.

[9] Baxter, M. and M. Crucini (1993) "Explaining Saving-Investment Correlations", *American Economic Review*, 83, 416-436.

[10] Baxter, M. and R. King (1993) "Fiscal Policy in General Equilibrium", *American Economic Review*, vol.83 (3), 315-334.

[11] Brock, W.A. and L. Mirman (1972) "Optimal Economic Growth and Uncertainty: The Discounted Case", *Journal of Economic Theory*, 479-513.

[12] Canova, F. (1994) " Statistical Inference in Calibrated Models", *Journal of Applied Econometrics*, 9, S123-S144.

[13] Canova, F. (1995) "Sensitivity Analysis and Model Evaluation in Simulated Dynamic General Equilibrium Economies", *International Economic Review*, 36, 477-501.

[14] Canova, F. (1997) "Detrending and Business Cycle Facts", *Journal of Monetary Economics*, forthcoming.

[15] Canova, F. and G. De Nicoló (1995), "The Equity Premium and the Risk Free Rate: A Cross Country, Cross Maturity Examination", CEPR working paper 1119.

[16] Canova, F., Finn, M. and A. Pagan (1994), "Evaluating a Real Business Cycle Model", in C. Hargreaves (ed.), *Nonstationary Time Series Analyses and Cointegration*, Oxford, UK: Oxford University Press.

[17] Canova, F. and E. Ortega (1996), "Testing Calibrated General Equilibrium Models", forthcoming in Mariano, R., Schuermann, T. and M. Weeks (eds.), *Simulation Based Inference in Econometrics: Methods and Applications*, Cambridge: Cambridge University Press.

[18] Cecchetti, S.G., Lam, P. and N. Mark (1993), "The Equity Premium and the Risk-free Rate: Matching the Moments", *Journal of Monetary Economics*, 31, 21-45.

[19] Cho, J. and T. Cooley (1994), "Employment and Hours over the Business Cycle", *Journal of Economic Dynamics and Control*, 18, 411-432.

[20] Cho, J. and T. Cooley (1995), "The Business Cycle with Nominal Contracts", in Cooley, T. (ed.), *Frontiers of Business Cycle Research*, Princeton, US: Princeton University Press.

[21] Christiano, L.J. and M. Eichenbaum (1992), " Current Business Cycle Theories and Aggregate Labor Market Fluctuations", *American Economic Review*, 82, 430-450.

[22] Christiano, L.J. and W. Den Haan (1995), "Small Sample Properties of GMM for Business Cycle Analysis", Federal Reserve Bank of Minneapolis Staff Reports 199.

[23] Cogley, T. and J.M. Nason (1994), "Testing the Implications of Long-run Neutrality for Monetary Business Cycle Models", *Journal of Applied Econometrics*, 9, S37-S70.

[24] Coleman, W. (1989) "An Algorithm to Solve Dynamic Models", Board of Governors of the Federal Reserve System, International Finance Division, Discussion Paper no. 351.

[25] Danthine, J.P. and J.B. Donaldson (1992), "Non-Walrasian Economies", Cahiers de Recherche Economique, Université de Lausanne, No.9301.

[26] Danthine, J.P. and J.B. Donaldson (1993), "Methodological and Empirical Issues in Real Business Cycle Theory", *European Economic Review*, No.37.

[27] DeJong, D., Ingram, B. and C. Whiteman (1996), "A Bayesian Approach to Calibration", *Journal of Business and Economic Statistics*, 14, 1-9.

[28] Den Haan, W. and A. Marcet (1994), "Accuracy in Simulations", *Review of Economic Studies*, 61, 3-17.

[29] Diebold, F., Ohanian, L. and J. Berkowitz (1995), "Dynamic Equilibrium Economies: A Framework for Comparing Models and Data", NBER Technical Working Paper No.174.

[30] Duffie, D. and K. Singleton (1993), "Simulated Moments Estimation of Markov Models of Asset Prices", *Econometrica*, 61, 929-950.

[31] Eberwein, C. and T. Kollintzas (1995), "A Dynamic Model of Bargaining in a Unionized Firm with Irreversible Investment", *Annales d'Economie et de Statistique*, 37-38, 91-116.

[32] Fève, P. and F. Langot (1994), "The RBC Models through Statistical Inference: An Application with French Data", *Journal of Applied Econometrics*, 9, S11-S37.

[33] Gali, J. (1994), "Monopolistic Competition, Business Cycles and the Composition of Aggregate Demand?", *Journal of Economic Theory*, 63, 73-96.

[34] Gali, J. (1995), "Real Business Cycles with Involuntary Unemployment", CEPR Discussion Paper No.1206.

[35] Geweke, J. (1989), " Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317-1339.

[36] Genest, C. and M. Zidak (1986) "Combining Probability Distributions: A Critique and an Annotated Bibliography", *Statistical Science*, 1, 114-148.

[37] Granger, C.W.J. (1964), *Spectral Analysis of Economic Time Series*, Princeton University Press.

[38] Gregory, A. and G. Smith (1989), " Calibration as Estimation", *Econometric Reviews*, 9(1), 57-89.

[39] Gregory, A. and G. Smith (1991), " Calibration as Testing: Inference in Simulated Macro Models", *Journal of Business and Economic Statistics*, 9(3), 293-303.

[40] Gregory, A. and G. Smith (1993), " Calibration in Macroeconomics", in Maddala, G.S. (ed.), *Handbook of Statistics*, vol. 11, Amsterdam, North Holland.

[41] Hannan, E.J. (1970), *Multiple Time Series*, John Wiley and Sons.

[42] Hansen, G. (1985), "Indivisible Labor and the Business Cycle", *Journal of Monetary Economics*, 16 (3), 309-327.

[43] Hansen, L. (1982), " Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, 50, 1029-1054.

[44] Hansen, L. and R. Jagannathan (1991), "Implications of Security Market Data for Models of Dynamic Economies", *Journal of Political Economy*, 99, 225-262.

[45] Hansen, L. and T.J. Sargent (1979), "Formulating and Estimating Dynamic Linear Rational Expectations Models", *Journal of Economic Dynamic and Control*, 2, 7-46.

[46] *Journal of Business and Economic Statistics*, January 1990.

[47] King, R., and C. Plosser (1994), " Real Business Cycles and the test of the Adelmans" , *Journal of Monetary Economics*, 33, 405-438.

[48] King, R., Plosser, C. and S. Rebelo (1988), " Production, Growth and Business Cycles: I and II", *Journal of Monetary Economics*, 21, 195-232 and 309-342.

[49] King, R., Plosser, C. and S. Rebelo (1990), " Production, Growth and Business Cycles: Technical Appendix", mimeo, University of Rochester.

[50] Kim, K. and A. Pagan (1994) "The Econometric Analysis of Calibrated Macroeconomic Models", forthcoming, in Pesaran, H. and M. Wickens, eds., *Handbook of Applied Econometrics*, vol.I, London: Blackwell Press.

[51] Kollintzas, T. (1992)," Comment to J.P. Danthine: Calibrated Macroeconomic Models: What and What for", manuscript, Athens University.

[52] Kuhn, T. (1970), *The Structure of Scientific Revolution*, Chicago, Il.: Chicago University Press.

[53] Kydland, F. (1992) " On the Econometrics of World Business Cycles", *European Economic Review*, 36, 476-482.

[54] Kydland, F. and E. Prescott (1982), "Time To Build and Aggregate Fluctuations", *Econometrica*, 50, 1345-1370 .

[55] Kydland, F. and E. Prescott (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles", *The Scandinavian Journal of Economics*, 93(2), 161-178.

[56] Lee, B.S. and B. Ingram (1991), "Simulation Estimators of Time Series Models", *Journal of Econometrics*, 47(2/3), 197-206.

[57] Lucas, R.E., Jr. (1976), "Econometric Policy Evaluation: A Critique", in Brunner, K. and A. Meltzer (eds.), *Carnegie-Rochester Series on Public Policy*, North-Holland, vol.1, 19-46.

[58] Lucas, R.E., Jr. (1980), "Methods and Problems in Business Cycle Theory", *Journal of Money, Credit and Banking*, vol.12, 696-715.

[59] Lucas, R.E., Jr. and T.J. Sargent (1981), *Rational Expectations and Econometric Practice*, Minneapolis: The University of Minnesota Press.

[60] Marcet, A. (1992) "Solving Nonlinear Stochastic Models by Parametrizing Expectations: An Application to Asset Pricing with Production", Universitat Pompeu Fabra, working paper 5.

[61] Marcet, A. (1994), "Simulation Analysis of Stochastic Dynamic Models: Applications to Theory and Econometrics" in Sims, C. (ed.), *Advances in Econometrics. Sixth World Congress of the Econometric Society*, Cambridge: Cambridge University Press.

[62] Mehra, R. and E. Prescott (1985), "The Equity Premium: A Puzzle", *Journal of Monetary Economics*, 15, 145-162.

[63] McGrattan, E. , Rogerson, B. and R. Wright (1993), "Estimating the Stochastic Growth Model with Household Production", Federal Reserve Bank of Minneapolis, manuscript.

[64] Monfardini, C. (1995), "Simulation-Based Encompassing for Non-nested Models: A Monte Carlo Study of Alternative Simulated Cox Test Statistics", European University Institute WP ECO No. 95/37.

[65] Niederreiter, H. (1988), "Quasi Monte Carlo Methods for Multidimensional Numerical Integration", *International Series of Numerical Mathematics*, 85, 157-171.

[66] Newey, W. and K. West (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-708.

[67] Novales, A. (1990), "Solving Nonlinear Rational Expectations Models: A Stochastic Equilibrium Model of Interest Rates", *Econometrica*, 58, 93-111.

[68] Ortega, E. (1995), "Assessing and Comparing Multivariate Dynamic Models", manuscript, European University Institute.

[69] Pagan, A. (1994), "Calibration and Econometric Research: An Overview", *Journal of Applied Econometrics*, 9, S1-S10.

[70] Pagan, A. and Shannon (1985), "Sensitivity Analysis for Linearized Computable General Equilibrium Models", in J.Piggott and J. Whalley (eds.) *New Developments in Applied General Equilibrium Analysis*, Cambridge: Cambridge University Press.

[71] Pesaran, H. and R. Smith (1992), "The Interaction between Theory and Observation in Economics", University of Cambridge, manuscript.

[72] Priestley, M.B. (1981), *Spectral Analysis and Time Series*, Academic Press.

[73] Rotemberg, J. and M. Woodford (1991), " Markups and the Business Cycle", in Blanchard, O.J. and S. Fisher (eds.) *NBER Macroeconomics Annual 1991*, Cambridge: MIT Press.

[74] Sargent, T. (1979), *Macroeconomic Theory*, New York: Academic Press.

[75] Sargent, T. (1987), *Dynamic Macroeconomic Theory*, Cambridge, Ma: Harvard University Press.

[76] Showen, J. and J. Whalley (1984), "Applied General Equilibrium Models of Taxation and International Trade: An Introduction and Survey", *Journal of Economic Literature*, 22, 1007-1051.

[77] Simkins, S.P. (1994), "Do Real Business Cycle Models Really Exhibit Business Cycle Behavior?", *Journal of Monetary Economics*, 33, 381-404.

[78] Smith, T. (1993) "Estimating Nonlinear Time Series Models Using Simulated VAR", *Journal of Applied Econometrics*, 8, S63-S84.

[79] Söderlind, P. (1994), "Cyclical Properties of a Real Business Cycle Model", *Journal of Applied Econometrics*, 9, S113-S122.

[80] Stadler, G.W. (1994), "Real Business Cycles", *Journal of Economic Literature*, vol.XXXII, 1750-1783.

[81] Summers, L. (1991), "Scientific Illusion in Empirical Macroeconomics", *Scandinavian Journal of Economics*, 93(2), 129-148.

[82] Tauchen, G. and R. Hussey (1991) "Quadrature Based Methods for obtaining Approximate Solutions to Integral Equations of Nonlinear Asset Pricing Models", *Econometrica*, 59, 371-397.

[83] Ubide, A.J. (1995), "On International Business Cycles", Ph.D. Dissertation, European University Institute.

[84] Watson, M. (1993), "Measures of Fit for Calibrated Models", *Journal of Political Economy*, 101, 1011-1041.

[85] Wolf, F. (1986) *Meta-Analysis: Quantitative Methods for Research Synthesis*, Beverly Hill, Ca.: Sage Publishers.